

Educating future scientists

Instead of helplessly pondering a new trove of data for guidance on how to improve science education, researchers should better exploit existing mechanisms for helping out at their local schools.

A quarter of a million 15-year-olds from around the world last year performed hours of special tests on their reading, mathematical and scientific literacy, in a study organized by the Organisation for Economic Co-operation and Development (OECD). The first results of this Programme for International Student Assessment (PISA) have just been released — and make moderately interesting reading.

Korea sits at the top of the science class, closely followed by Japan, Finland and the United Kingdom. Given its wealth, the United States (which has the widest gap between its best and worst students) should certainly do better than its mid-table position. Germany, surprisingly, performs the least impressively of the major scientific nations, with results below the overall international average. The science results closely mirror those in reading and mathematics, and it might be tempting to draw instant conclusions from them: some in Germany have already done so. But caution is appropriate in interpreting such studies, which really provide more questions than answers for policy-makers.

The most noted recent international comparison of school performance was the Third International Mathematics and Science Study (TIMSS), which was completed in 1999. But the questions in TIMSS sought, in the main, to measure students' grasp of the facts contained in school curricula. PISA, in contrast, sought to measure their ability to apply knowledge.

This attempt to measure the evaluation and interpretation skills of children should be welcomed. But it would be a mistake to assume that the scientific futures of countries at the summit of this particular PISA tower are necessarily going to be rosier than those of nations that fared less well. Just because a student is good at a subject, there is no guarantee that he or she will continue to study it. As generations of teachers will confirm, there is often little correlation

between academic ability, which the study sought to measure, and enthusiasm.

Indeed, many of those directly involved in school education speak of an increasing reluctance among students to continue studying science once it ceases to be a compulsory subject. The reasons for this have been well rehearsed. Science faces stiffer competition from other subject areas than previously, both in the struggle for attention and in the promise it offers of assisting a successful career. It is also relatively content-heavy — however it is packaged — and even the most skilful teacher may struggle to make some of its more esoteric aspects relevant to the world of the average 13-year-old.

The scientific community has a stake in these problems, but it is one tinged with paradox. A more accessible — and perhaps less challenging — curriculum might well encourage wider participation and a larger pool of students willing to continue studying science. But such a curriculum will not best help the very brightest students whose interest in science might lead them into a career in it. The argument about which direction to push the curriculum will doubtless continue. As it does so, scientists can do more than grumble from the sidelines about declining standards.

Many scientific organizations already take a strong interest in science education in schools. The US National Science Foundation — whose charter includes supports for education as well as research — now asks all grant applicants to explain what their research proposal will do for education, and this can include school education. In Britain, a programme called the Pupil Researcher Initiative, funded by the research councils, offers school students the chance to regularly hear from active researchers. A wealth of other such schemes operate around the world. Now is the time for researchers — some of whom are reluctant even to engage properly in undergraduate education — to seek these out and participate in them. ■

Europe must unite to preserve its heritage

The European Commission should not lose its enthusiasm for supporting research into the preservation of cultural heritage.

Relatively speaking, it's a discipline that doesn't cost much — and its value is beyond measure. Yet research into the conservation of Europe's vast cultural heritage has few real champions. Its best friend is Italy, the only country in the European Union (EU) that provides serious national funding for research into the preservation of monuments, ancient ruins and archaeological sites.

But even Italians concede that the Leaning Tower of Pisa, for example, is no more representative of Europe's heritage than, say, the historic centre of a village in Norway. Conservation efforts should be truly European in scope, not only to acknowledge Europe's cultural unity — which is supposedly central to the EU charter — but for technical reasons, too. It is widely accepted, for example, that the standardization of procedures for building preservation will save money and help to efficiently share out the limited pool of available expertise.

For 15 years, the European Commission has shown itself sensitive

to these issues. Its approach recognized the fact that national efforts outside Italy tend to be narrowly focused and piecemeal. Unfortunately, however, the commission's proposal for the next Framework research programme, which starts in 2003, makes no mention of the conservation of cultural heritage. National governments, it implies, should handle the problem themselves. This attitude is short-sighted.

The list of research needed for heritage conservation is long, and underlines the need for continued support from the EU. Thresholds for air quality should be set, for example, with building conservation goals in mind. Research is badly needed into the interactions between ancient building materials and the modern alternatives used in restoration. Movement of the water table, which is affected by global warming and controls the destiny of buried archaeological ruins, needs to be better understood. Italy knows it cannot afford to let Venice float away to sea, but the rest of Europe needs to ensure that other aspects of its valuable heritage are preserved for future generations. ■





Expense account
Japan's offer to pay
for trip unnerves
Nobel officials
p676



Testing times
Trial raises fresh
question mark over
gene therapy
p677



Paper auction
Wellcome Trust
secures Crick archive
for London library
p678



Star wars
Space-station
partners rail against
NASA proposal
p679

Bioweapons treaty in disarray as US blocks plans for verification

Declan Butler, Paris

The United States is facing bitter recriminations over its tactics at an international meeting on biological weapons in Geneva. The talks to review the 1972 Biological Weapons Convention (BWC) collapsed on 7 December without agreement.

A unanimous vote to adjourn the conference until next November followed a move by the United States to block new means for verifying compliance with the BWC's terms.

As the meeting finished, negotiators from European Union (EU) states openly condemned the US tactics — echoing the deep divisions that characterized international conferences on topics such as global warming before the 11 September terrorist attacks. The EU delegation was sufficiently annoyed that it refused to sit with its US allies during the closing session of the meeting.

Critics of the Bush administration say it has committed a sizeable diplomatic blunder, alienating its closest allies and also failing in its bid to stop the BWC parties from attempting fresh negotiations for a verification protocol in the future. But the administration believes that it succeeded in postponing the difficult issue of international agreement on bioweapons, without sustaining any real political damage.

The adjournment leaves the BWC in disarray. But some observers say that the enforced 'cooling off' period, together with the galvanization of opposition at the meeting to the Bush administration's position, may ultimately result in greater international impetus towards an expanded BWC agreement.

Less than two hours before the scheduled end of the conference, the United States tabled a demand that other countries were always going to reject: the formal abolition of a remit requiring parties to the BWC to agree on legally binding measures to ensure compliance. In July, the United States had said that it would veto a draft protocol allowing for inspections of facilities, arguing that this would not catch bioweapons proliferators, and would open up US biodefence establishments and biotechnology companies to espionage (see *Nature* **412**, 365; 2001).

From the outset, it was recognized that



Fighting talk: under-secretary of state John Bolton (right) raises US concerns in Geneva.

discussion of this remit at the meeting would probably result in the breakdown of the talks, which required unanimous consent to proceed. So parties representing the BWC's 144 signatory states had tacitly agreed to ignore the issue, negotiators say.

Instead, the talks sought consensus on two other issues. One was the nature of future BWC review meetings. To get around US opposition to formal discussion of verification protocols, the EU proposed ad hoc meetings alongside the five-year review conferences.

The other was the question of how to deal with signatories suspected of running covert, offensive biological weapons programmes. Initially, the United States demanded that five named states — including Iran and Iraq — “terminate their offensive biological weapons programmes and fully comply with their obligations”. This irked some parties because it ignored eight other nations — including Russia, China and Israel — that the United States acknowledges as having similar programmes.

On the last morning of the meeting, the United States offered a concession, asking instead that parties express “grave concern that compliance... has been subject to doubt in certain cases”. This was expected to win the meeting's support, negotiators say.

But with agreement in sight, the United States abandoned its position, saying that it would only agree to the compromise if the



convention's remit of establishing a verification protocol was formally terminated. Deadlocked on that last-minute proposal, the meeting adjourned, shelving all discussions until 2002.

“We have a real profound difference with the United States as to how you strengthen the BWC,” says one angry UK government official, adding that “Britain still believes that multilateral arms control has a future”. Elisa Harris, a bioweapons adviser to the Clinton administration, describes the US action as “not just irresponsible, but shocking”.

Michael Powers, an analyst at the Chemical and Biological Arms Control Institute in Washington, contests this viewpoint. He argues that the United States is now in the “vanguard” of a more pragmatic international effort towards arms control, away from fixed multilateral treaties and towards more flexible approaches.

Whatever happens, over the next year the United States will need to embrace some sort of international agreement on bioweapons, predicts Matthew Meselson, a biological-weapons specialist at Harvard University. Meselson favours an alternative approach that makes production or use of such weapons a criminal offence under international law, like torture or aircraft hijacking. ■

AP

Nobel officials recoil from expenses offer

David Cyranoski, Tokyo

An offer by a Japanese government organization to pay for Nobel Foundation officials to visit prize-hungry Japan is ruffling feathers in Stockholm.

The Science Council of Japan says it will pay for Nobel laureates and representatives — including the director and secretary of each of the Nobel Foundation's three scientific committees — to travel to Tokyo next March for a forum celebrating 100 years of the Nobel prizes.

But one of the invited officials says the offer to cover the trip's expenses is unnecessary and inappropriate — especially at a time when Japan has set a national target of winning 30 Nobel prizes in the next 50 years.

"The invitations pose an ethical problem for me as there is such an outspoken Japanese policy to acquire Nobel prizes," says Anders Bárány, a physicist at Stockholm University and secretary of the physics committee. "The Nobel Foundation has enough money to pick up the bill."

The government's Nobel target has been attacked by some Japanese academics as inappropriate (see *Nature* 413, 562; 2001). But others have defended it as a harmless goal that will spur researchers and help to ensure political support for science in Japan.



Role model: Ryoji Noyori won a Nobel prize this year, but Japan wants 30 more like him by 2050.

The March forum will coincide with the opening of a touring exhibition on the Nobels, for which Japan is the first stop. "In other international events, it has been customary to cover invitees' costs," says Kiyoshi Kurokawa, vice-president of the science council. "This isn't special treatment." But he adds that the council "will wait to see how the Nobel Foundation wants to handle expenses" before it actually pays them.

Not all Nobel Foundation officials are concerned by the proposed arrangement. "For travel and hotel expenses to be covered by the inviting organizers seems to be standard procedure," says Michael Sohlman, executive director of the Nobel Foundation.

Bárány says he supports the forum itself, but fears a repeat of the controversy that attached itself to the 1986 award of the medicine prize to Rita Levi-Montalcini of Italy's Institute of Cell Biology in Rome. It was later alleged in Swedish newspapers that a Nobel committee member had accepted an expenses-paid trip from a drug company that wanted Levi-Montalcini to get the prize. "Investigations showed that the Nobel committee member just failed to realize that this kind of question can come up," says Bárány, adding that this was "silly, as each committee has money for this kind of travel anyway".

But there are no clear guidelines for what committee members can or cannot accept in terms of invitations. Bárány says he wants the Nobel Foundation to address this issue quickly. But Sohlman does not think it is a problem. "We have been relying on the common sense of the members of the Nobel institutions for a hundred years now," he says, "and have reason to continue to do so." ■

Mathematicians poised for major funding boost

Erica Klarreich

Increased support for mathematics research is set to form a central plank of the 2003 budget for the US National Science Foundation (NSF) to be proposed by President George W. Bush in February.

Mathematicians hope that the increase will allow the agency to deliver larger grants. Compared with other disciplines, where the median NSF grant is \$80,000 per year, mathematics' median is only about \$35,000

— described as "a disgraceful number" by Philippe Tondeur, head of the NSF's mathematics division.

Back in October 2000, NSF director Rita Colwell called for the mathematics budget — \$121 million in 2001 — to quadruple by 2007 (see *Nature* 407, 931; 2000). But the 2002 budget will only increase the subject's allocation by about \$20 million, far short of Colwell's goal. Tondeur says that he hopes the 2003 figure will be "on a higher scale" than that for 2002. Senior officials in the agency say he will not be disappointed.

The government cannot release details of the budget in advance, but Sam Rankin, associate executive director of the American Mathematical Society (AMS), says that mathematicians are looking for a division budget of as much as \$200 million.

The money would pay for new interdisciplinary projects linking mathematics to biology, computer science and geosciences, Tondeur says. It would also allow the division to expand its programme of Grants for Vertical Integration of Research and Education, launched in 1999. This programme, which already funds 31 departments, encourages interaction

between senior faculty, postdoctoral researchers, graduate students and undergraduates.

US mathematicians complain that funding for their discipline has been declining since the 1970s as defence agencies, which used to fund much of their work, became less supportive of basic research. The NSF now accounts for nearly three-quarters of federal funding of mathematics. But only about 1,500 mathematicians — a small fraction of academic mathematicians — get NSF grants each year, according to Rankin.

"The backbone of mathematics is the individual investigator, and the system has not kept pace with the number of strong investigators in recent years," says David Eisenbud, director of the Mathematical Sciences Research Institute in Berkeley, California.

The new initiative should create a very different picture, Tondeur says. "It's going to be a really adventurous period for the mathematical sciences," he says. "There's a growing perception that mathematics is an enabling discipline for science, technology and engineering." ■



Trial halted after gene shows up in semen

Nell Boyce, Washington

A patient in a gene-therapy trial has had the virus used to transfer the gene show up in his semen. The finding highlights the danger that gene therapy may inadvertently modify the genetic make-up of a patient's children.

The discovery was made in a trial at the Children's Hospital of Philadelphia, conducted in collaboration with researchers at Stanford University Medical Center in California. Officials at the Food and Drug Administration (FDA) halted the trial.

But the Recombinant DNA Advisory Committee (RAC), which advises the FDA on the safety and ethics of such experiments, announced on 6 December that the trial could continue, subject to special monitoring. The FDA is expected to follow this advice.

Gene therapy often uses viruses to transport therapeutic genes into cells. But these viruses can become widely dispersed in the body, and animal studies have shown that they can end up in the ovaries and testes. The FDA, which regulates clinical trials in the United States, has been concerned that the therapeutic genes might be passed on to future generations. In 1998 it ruled that many human gene trials that ran the risk of contaminating eggs or sperm could enrol only sterile patients. The rule exempted trial participants whose lives were in immediate danger, but covered trials for treatable diseases.

Some scientists and patients' groups have argued that this impedes research while protecting against what they consider to be a very small risk. No animal studies have shown gene transmission to the next generation; researchers suspect that evolution has devised ways of protecting the genetic cargo of eggs and sperm from viral onslaught.

As a precaution, the FDA requires investigators to test the semen of men taking part in trials where preclinical animal research has detected the gene in the gonads. That was the case in the Stanford study, which put the human gene for blood-clotting factor IX into the livers of men with severe haemophilia.

Mark Kay, one of the geneticists leading the study, says that the 63-year-old patient received a low dose of a virus carrying the gene on 13 August. Tests on his semen subsequently detected gene sequences from the virus. This result persisted for seven weeks.

But it is not known whether the virus actually entered the sperm, as semen also contains other cell types such as white blood cells. To investigate this possibility, Kay's colleagues separated a semen sample into its constituent cell types and identified the cells containing virus gene fragments. The sperm itself tested negative.

The FDA has said that no additional patient can be enrolled in the trial until each patient's semen tests negative for three



Hopes and fears: could a therapeutic gene be inadvertently passed on to future generations?

months — the time it takes for sperm to mature. But Kay's team plans to enrol nine patients in the trial one after the other, so this

could extend it for up to five years, Kay asserts. He says the risk of children being affected is reduced because patients are advised to use condoms. "If you look at the realm of other risks that people take in drug studies, this one is relatively small," he says.

The RAC has agreed, recommending that the trial be allowed to continue, provided researchers conduct frequent tests on isolated cell types from patients' semen samples.

The ethical issues will deepen only if researchers find that an effective gene therapy also shows up in sperm. "We're dealing here with the grey area in which you may knowingly allow something but not deliberately do it," says Ruth Macklin, a bioethicist at the Albert Einstein College of Medicine in New York, and a member of the RAC. ■

Need for vaccine stocks questioned

Erika Check, Washington

US biodefence experts are divided this week over a research paper that calls into question a government decision to spend hundreds of millions of dollars on smallpox vaccine.

The paper's authors say that 40 million doses of vaccine is enough to protect the country from a terrorist attack. Published in the current issue of *Emerging Infectious Diseases* (7, 959–969; 2001), the paper was written by Martin Meltzer, an economist at the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, with J. Donald Millar, who led the CDC's effort to eradicate smallpox, and two CDC biologists.

Health secretary Tommy Thompson announced on 28 November that the United States will buy 155 million doses of vaccine from Acambis, a British firm, for \$428 million, bringing the total effective US stockpile to 286 million doses.

Government scientists defended the purchase. Donald Henderson, who led the global effort to eradicate smallpox and now advises Thompson on biodefence, suggests the authors were influenced by the CDC's earlier decision to hold 40 million doses.

"Amazingly, they reached the conclusion that this indeed was the correct number to purchase," Henderson says.

Other researchers dispute the paper's assumption that every patient would infect an average of three others. Steve Leach and Raymond Gani, of Britain's biological weapons laboratory at Porton Down near Salisbury, report in *Nature* this week (pages 748–751) that each smallpox patient is likely to infect from 3.5 to 6 others, and even more initially, because people now lack immunity

to the disease, which was eradicated worldwide in 1976.

But Meltzer says his team's work was rigorous. "As researchers, my colleagues and I try to approach the analysis of each and every problem in a neutral fashion, employing scientific procedures," he says.

There is concern that stockpiling vaccine is a waste of scarce public-health resources, as a move to vaccinate the entire US population would result in hundreds or thousands of deaths from side effects. But some officials who agree the stockpile may not be needed argue that it still reassures the public. "It may be politically necessary and prudent to make sure that we allow a large safety margin," says Stephen Morse, director of the Center for Public Health Preparedness at Columbia University in New York. ■



Helpline: queues form in the Bronx, New York, to receive smallpox vaccine in 1947.

Wellcome bid sees Crick archive return home

Alison Abbott and Rex Dalton

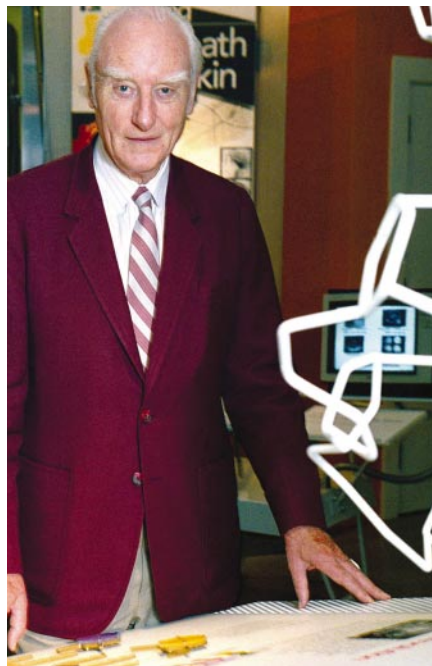
Francis Crick, co-discoverer of the double-helical structure of DNA with James Watson, has sold his scientific archive to Britain's Wellcome Trust for just under £2 million (US\$2.8 million).

The deal was struck after months of negotiations with at least one other bidder serving as a buyer for Jeremy Norman — a private collector based in Novato, California, who is accumulating a large archive of the history of molecular biology (see *Nature* **411**, 732–733; 2001).

The sum paid by the trust, which was bolstered by money from the UK national lottery fund, is understood to be the largest ever given for the archives of a living scientist. Crick's collection includes correspondence with other scientists, including Watson, together with his laboratory notebooks and drafts of articles and books.

The archive will be shipped to the trust's library in London from San Diego, where Crick, now 85, works part-time at the Salk Institute in La Jolla. He is donating a photocopy of the archive's contents to the University of California, San Diego.

Norman's representative, Al Seckel, a neuroscientist at the California Institute of Technology in Pasadena, has helped to acquire many other items for the Norman collection. Seckel had been trying to set up a centre for the history of molecular biology at his home institute, where Watson once



Record deal: Francis Crick's archive may be more expensive than that of any other living scientist.

worked, and hoped that Norman's collection could be included in it. Norman already owns papers from other eminent UK-based molecular biologists, including Maurice Wilkins, Max Perutz and Aaron Klug.

But many scientists, including Watson, have criticized privately held archives. Such

archives, they say, could restrict access to scholars and are prone to being broken up when the collector dies. Watson is currently compiling his own archive at the Cold Spring Harbor Laboratory in New York state, where he is director.

Norman has denied these possibilities, but publicity about his archive has increased the pressure on scientists to donate their papers to public institutions. Crick is understood to have accepted slightly less from Wellcome than Norman had offered.

Norman declined to comment on the negotiations. "The Wellcome library is a very appropriate depository for Francis Crick's papers," he says.

But Seckel feels that the Wellcome arrangement has stymied his vision of a unified archive. "It would have been better to have all the relevant molecular-biology papers together, rather than segregate Crick's in a library that only collects English scientists' papers," he says. "This is not the way science operates — it is not nationalistic."

Meanwhile, the prices being sought for historically important scientific documents keep growing. For example, a signed single-page proof from the 25 April 1953 issue of *Nature* is currently on sale by a rare-book dealer at an asking price of \$43,500. The page includes the famous Watson and Crick paper, as well as others by Wilkins and by Rosalind Franklin, whose X-ray crystallography helped to elucidate the structure of DNA. ■

WELLCOME TRUST

Jodrell Bank survives shake-up of UK astronomy

David Adam, London

Faced with tough astronomy funding choices after deciding to join the European Southern Observatory (ESO), Britain has elected to retain its ageing Jodrell Bank telescopes and axe its commitment to several overseas observatories.

At a council meeting on 5 December, the Particle Physics and Astronomy Research Council (PPARC), which funds most of Britain's astronomy, agreed reductions of some £5 million (US\$7.2 million) a year in its existing ground-based programmes to pay for its subscription to the Chile-based ESO project.

The cuts will affect observatories in Australia and Hawaii, as well as a cluster of optical telescopes at La Palma in the Canary Isles, where PPARC is pulling the plug on the Jacobus Kapteyn telescope and phasing out support for the Isaac Newton telescope.

But the biggest loser will be the Anglo-Australian observatory at Siding Spring in

southern Australia, for which funding will be phased out altogether by 2007.

"With a limited budget there will be inevitable reductions to certain facilities within the existing programme," says Ian Halliday, chief executive of PPARC.

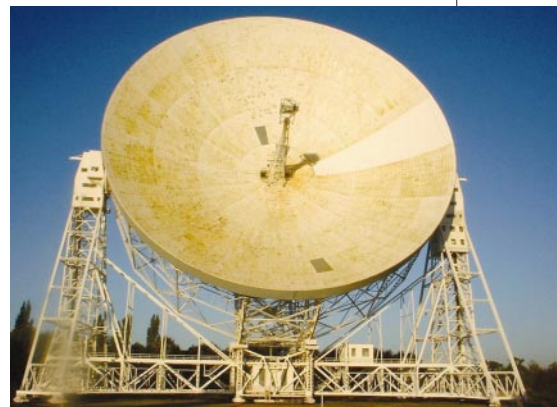
The decision secures the immediate future of the MERLIN array of radio telescopes, including the 76-metre Lovell Telescope at Jodrell Bank in Cheshire, near Manchester. The network will now be refurbished with fibre-optic cable.

"The upgrade to MERLIN will ensure that our own national facility, centred around the Lovell telescope in the northwest, will continue to deliver world-class science," says Martin Rees, the Astronomer Royal.

It will also help to maintain the discipline's profile in Britain — the dish at Jodrell Bank is a national landmark, and is the astronomy programme's most powerful symbol.

Joining the ESO will give British

astronomers access to the Very Large Telescope in Chile, as well as greater involvement in the Atacama Large Millimeter Array (ALMA), a network of 64 twelve-metre telescopes planned for construction there. ■



Star turn: the main Jodrell Bank telescope faces an upgrade rather than the axe.

Partners' anger mounts over NASA plans for space station



All dressed up, but nowhere to go? Space station cutbacks mean fewer European astronauts may fly.

Tony Reichhardt, Washington

The International Space Station's non-US partners arrived in Washington last week to vent their mounting anger at NASA's plan to trim back the project unilaterally. But it remains unclear what — if anything — they can do about it.

Representatives from Europe, Canada, Japan and Russia told NASA's advisory council that plans for the slimmed-down station now backed by the White House would not fulfil NASA's obligations under the international agreement to build and operate the station. The three-person, interim plan was sanctioned last month by a panel of US experts who reviewed the project (see *Nature* **414**, 136; 2001).

Space officials from Europe and Canada have written to the US State Department complaining about the plan and stating that the international agreement on the station calls for any major changes in the design to be agreed by all the partners. At his senate confirmation hearing on 7 December, Sean O'Keefe, who has been nominated by President Bush to lead NASA, confessed to a "limited understanding" of NASA's legal obligations under the agreement, and said he would consult with the State Department on the matter.

But the partners told the quarterly meeting of the advisory council that the United States cannot renege on its agreement to provide living quarters and a "lifeboat" for a six- or seven-person crew — despite the fact that NASA faces cost overruns of at least \$5 billion on the project.

"We have had our financial problems, too," said Jörg Feustel-Büechl, director of

manned spaceflight programmes for the European Space Agency. "We solved this problem at home, and we expect the same from you."

He added that European science ministers had "serious doubts" about continuing with the project after reports of US cutbacks — and that even a two-year delay before committing to the full station would be strongly opposed by Europe. A crew of three would have virtually no time to conduct research, and far fewer non-US astronauts could fly.

But with little real leverage to force NASA's hand, the partners fell back on the familiar line that the situation will damage collaboration on unspecified future projects. "There is much more at stake than just this programme," warned Feustel-Büechl.

Although the international agreement on the station calls for NASA to provide the living quarters and lifeboat, it also says that partners' obligations to the project depend on them actually obtaining funds from their respective governments.

The partners said they want a clear commitment from the US government to the full station. But a discussion among the advisory council members showed just how hard that might be to obtain. Some members argued that such a commitment, regardless of the final cost, would relieve any pressure on NASA to reform its management of the project.

At his confirmation hearing, meanwhile, O'Keefe said that NASA will have to "get our house in order" before committing to the full station.

Europe's Mars mission to pay out for Beagle lander

Sally Goodman, Paris

The European Space Agency (ESA) has been forced to allocate 36 million euros (US\$32 million) to a UK-designed lander to ensure that it is ready for launch as part of the Mars Express mission in 2003.

But members of ESA's Scientific Programme Committee, who agreed the extra funds on 4 December, said that no more money will be made available for the lander, which is called Beagle 2. They added that Mars Express — Europe's first mission to the red planet — will fly with or without the lander.

Beagle 2, which will search for organic material on Mars, was a late addition to Mars Express and costs for it were not included in the original budget of 150 million euros, which was finalized in 1996. The Beagle 2 consortium, led by Colin Pillinger of the Open University in Milton Keynes, England, was given the go-ahead to join the mission in 1998, provided it could raise sufficient funds. The consortium has since raised enough money from the British government, industry and private sponsorship deals to build the lander and its instruments.

But the inclusion of Beagle 2 is adding costs to the mission itself, such as those associated with keeping the lander in sterile conditions during the journey to Mars. ESA officials say they had expected to contribute something towards these, but that technical problems have increased the amount the agency will need to spend.

"We are doing everything we can to support Beagle 2," says Marcello Coradini, ESA's programme coordinator for Solar System missions.

Beagle 2's soft-landing system, which is made up of three gas-filled airbags, was responsible for the latest hitch. One or more of the airbags burst on inflation during tests last month at NASA's Plum Brook facility in Sandusky, Ohio. Such setbacks are delaying the lander's incorporation into the project, raising ESA's costs.

Pillinger remains confident that the project is on track, and points out that NASA needed extensive testing for the airbag system on Pathfinder, a craft that successfully landed on Mars in 1997. "We have a five-month contingency in our testing schedule for the airbags and plenty of time to adjust the design," he says.

Anger spreads over White House plans for Smithsonian

Washington Researchers and scientific advisers at the Smithsonian Institution are up in arms over reports that the White House wants to transfer the museum complex's best-known research facilities to the National Science Foundation (NSF).

The facilities, including the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, and the Smithsonian Tropical Research Institute in Panama, would be shifted to the NSF in 2003 under a plan hatched by the powerful White House Office of Management and Budget.

Researchers such as Melinda Zeder, an anthropologist at the Smithsonian's National Museum of Natural History in Washington, fear that the proposed budget of \$35 million would not cover the institutes' running costs.

Jerry Sabloff of the University of Pennsylvania, who chairs the Smithsonian's recently appointed science commission, says the topic will be hotly debated at the commission's next meeting on 13 December.

Sabloff, who wants to ensure that the museum complex continues to perform first-class research, says he hopes that further debate will lead to the proposal being shelved before President Bush's 2003 budget is announced in February.

Survivor brings personal experience to cancer post

Washington President George W. Bush has nominated Andrew von Eschenbach, a cancer survivor and researcher at the M. D. Anderson Cancer Center at the University of Texas in Houston, as director of the National Cancer Institute (NCI), part of the National Institutes of Health (NIH).

Von Eschenbach, whose nomination is expected to be quickly confirmed by the US Senate, currently runs the Genitourinary Cancer Center and the prostate cancer research programme at M. D. Anderson. He was also president-elect of the American Cancer Society, and was a founder of the

National Dialogue on Cancer, which aims to encourage investment in cancer organizations. John Mendelsohn, president of M. D. Anderson, called von Eschenbach a "superb clinician and a worldwide leader in prostate-cancer research".

Richard Klausner,

the NCI's last director, left in September to lead a new research organization, the Case Institute of Health, Science and Technology. Researchers are pleased that Klausner has been rapidly replaced, but are concerned that six other institutes at the NIH — as well as the NIH itself — remain without permanent leaders.

Two become one in battle against cancer

London Britain's two leading cancer charities are to merge to form a new organization with a combined annual expenditure of £130 million (US\$190 million). The Cancer Research Campaign (CRC) and the Imperial Cancer Research Fund (ICRF) were expected to announce their merger at a press conference on 11 December. The new charity will be known as Cancer Research UK.

The alliance has been widely expected following discussions that became public earlier this year (see *Nature* 409, 4; 2001). By combining the ICRF's focus on basic, molecular science with the more clinical approach of the CRC, the two charities aim to create a new body loosely modelled on the publicly funded US National Cancer Institute, which conducts both types of work in parallel.

Livermore lab director set to quit next year

San Diego Bruce Tarter, the director of the Lawrence Livermore National Laboratory in California, has announced that he will be leaving his post next year.

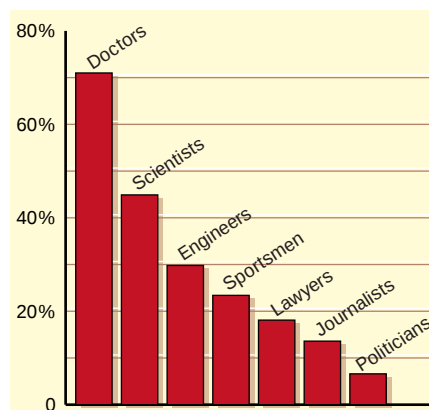
In a seven-year tenure, Tarter has ensured the future of the nuclear weapons laboratory by securing the construction of the National Ignition Facility there. But the project — a high-intensity laser chamber designed to ignite nuclear fusion — has been beset by large cost overruns and delays, and is not expected to open until 2008, overshooting its original opening time by some five years.

The University of California, which manages the laboratory for the US Department of Energy, says it has already started looking for Tarter's replacement. Tarter, a theoretical physicist, has been with the lab since 1967.

Scientists win respect but not public interest

Munich A new survey has revealed a mixed set of attitudes towards science among the European public. Europeans continue to trust scientists more than other professions, but less than half of all Europeans say they are interested in science.

Doctors, scientists and engineers took the top three spots when the public were asked which professions they held in the highest



A head of esteem? Many Europeans say they hold doctors and scientists in high regard.

esteem (see graph above). If a disaster struck in their neighbourhood, for example, most Europeans say they would prefer to turn to scientists, rather than journalists or government officials, for explanation.

But the Eurobarometer survey, published last week by the European Commission, revealed that more than 60% of those surveyed did not enjoy reading about science, and only 46% described themselves as interested in the subject. The commission had already earmarked 50 million euros (US\$44 million) to pay for initiatives towards the public understanding of science during its next Framework programme for research.

♦ <http://europa.eu.int/comm/research>

Europe loses its way over satellite navigation funding

Munich The future of Galileo, the proposed European satellite navigation system, has been thrown into doubt by a new report from consultancy firm PriceWaterhouseCoopers. The report questions the extent to which industry would be prepared to fund the project.

European Union (EU) transport ministers agreed in April to co-fund with industry the project's start-up costs of 3.4 billion euros (US\$3 billion). Galileo, which is due to become operational in 2008, would provide an alternative to the US-run Global Positioning System (see *Nature* 410, 853; 2001).

EU member states were expected to provide seed money for the system in their 2002 budgets, but Germany and Britain said on 7 December that they needed until at least March to consider the report, which was released last month. Contrary to EU plans, the study says that public money will be needed to run the project after 2008. Industry participation in the starting phase will also be less than expected.

Loyola de Palacio, the European transport commissioner, maintains that the European Commission will drop the project if member states do not make binding commitments by the end of this year.



Von Eschenbach: now heads cancer institute.

Which way to energy utopia?

The vehicles of the future will almost certainly be powered by hydrogen. But no one is sure exactly how to get drivers to kick their fossil-fuel habit. Mark Schrope weighs up the options.

Imagine a world in which everyone uses all the energy they want, yet dependence on oil, with its attendant smog and greenhouse-gas emissions, is a thing of the past. This utopia is plausible — many would say probable. It is one in which hydrogen, rather than fossil fuels, is central to our energy economy.

In the most ambitious vision, all the hydrogen we need would be liberated from water using renewable energy sources. Burning it, or using it in battery-like fuel cells that produce electricity, would allow the gas to power everything from buildings to cars. Fossil fuels would be almost totally removed from the energy equation.

Although this is seen by most as the best outcome, the way forward is far from clear. The basic technology needed already exists. But renewable energy sources are not mature enough to provide all the hydrogen we would need, so alternative methods are under consideration (see 'Cranking up the hydrogen flow', opposite). And it remains unclear how best to lure people away from conventional cars and into hydrogen-powered alternatives.

The transport sector is seen by experts as central to the hydrogen economy. It is the main driver of oil dependence — something the developed world wants to reduce, given the vulnerability of supplies to conflict and political disturbance in oil-exporting nations. Exhaust gases from vehicles also damage air quality, and transport as a whole is responsible for around a



It's a gas: a prototype hydrogen-powered vehicle from BMW refuels during testing.

third of all greenhouse-gas emissions.

Debate is now raging over which technologies should be used as stepping stones to a hydrogen future for vehicles. If the politicians, companies and environmentalists involved can reach a consensus, the prize is huge: cars and buses that emit nothing more than water vapour.

Burning issues

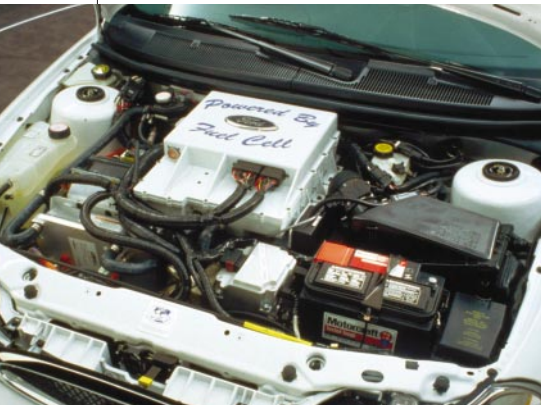
Vehicles could use hydrogen in a variety of ways. Some researchers favour the introduction of electric cars powered solely by fuel cells, which combine hydrogen and oxygen to produce electricity. Others say that conventional car engines can be converted to run on hydrogen with relatively minor modifications. Experts are also split over whether, as an interim step towards a full hydrogen economy, vehicles should initially use on-board equipment to extract hydrogen from fossil fuels.

Infrastructure issues play a big role in the debate over which approach should be taken. The lack of an existing system for storing and distributing hydrogen presents a dilemma. Car manufacturers do not want to sell

vehicles that people cannot fuel, and energy companies do not want to spend money developing a hydrogen distribution infrastructure when there are no hydrogen cars on the road. The equation becomes more complicated with fuel cells because they have yet to be produced in large numbers and their long-term reliability has not been proven.

This deadlock could be broken by 'reformers', which would allow hydrogen cars to run on fossil fuels. Reformers can break down the hydrocarbons in fossil fuels and so liberate hydrogen. Natural gas, for example, can be reformed by heating it together with water and a nickel-based catalyst. The result is a series of reactions whose products are carbon dioxide and hydrogen. Other fossil fuels, including petrol or gasoline, can be reformed in a similar way.

Hydrogen cars fitted with reformers would still run on petrol, but would reform it into hydrogen. Advocates of the technology say that this would give car companies the confidence to produce the vehicles, and so provide a fresh impetus for fuel-cell development. Several car manufacturers, including General Motors and DaimlerChrysler, are



Powerhouse: cars running on hydrogen fuel cells emit nothing more than water vapour.

now working with Ballard Power Systems, a fuel-cell producer based in Burnaby, near Vancouver, to develop vehicles that are powered by fuel cells fed by reformers.

But reformers still produce carbon dioxide, and for many environmentalists, this is enough to rule them out. "If you are going to move to a technology that allows you to have literally no pollution, why would you want to salvage the pollution in the process of switching?" asks Dan Becker, director of the global warming and energy programme at the Sierra Club, a San Francisco-based environmental organization. "It's like a nicotine patch that causes cancer."

Hydrogen vehicles with reformers are also technologically more complex and costly to build than straight fuel-cell cars, argues Robert Williams of the Center for Energy and Environmental Studies at Princeton University in New Jersey.

Reformed character

Some argue that these objections could be reduced by using methanol, rather than petrol, to power reformer vehicles. Reforming methanol, a low-mass alcohol that is liquid at room temperature, is more efficient than the process for petrol, as the hydrogen can be extracted at a lower temperature. It also produces fewer greenhouse-gas emissions. And because methanol can be produced from coal and natural gas, it could also reduce dependence on oil.

But methanol would need new infrastructure to distribute it. As a liquid, the changes required would be less than for hydrogen. A distribution system for ethanol, another liquid alcohol, is already in place in Brazil, where the fuel is produced from sugar cane. But, critics argue, developing a methanol distribution network would take us no closer to the hydrogen system that most agree is the ultimate goal.

Like petrol, methanol is poisonous. But it has only a slight odour and mixes easily with water, raising the risk that it could contaminate water supplies without people knowing. Another drawback of all reformer systems is that they currently require several minutes' heating before they can operate — far too long for drivers who expect to get in their cars and go instantly.

Given these problems, some experts argue that the most sensible route to a hydrogen economy is to tackle the infrastructure issue head on. With proper planning, says Williams, a hydrogen infrastructure could be created gradually and economically. Fleet vehicles, such as city buses, government vehicles and delivery trucks, would be the starting point. Because all the vehicles return to the same place every night, only one refuelling station would be needed for each fleet. Prototype projects of this type are currently under way in several European

Conventional internal combustion engines could be converted to burn hydrogen with fairly minor alterations.

countries and the United States.

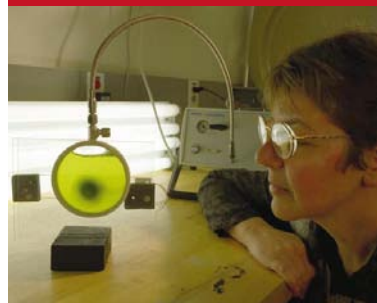
Williams suggests that governments could then encourage further dissemination of the technology by requiring vehicle manufacturers to sell a certain number of fuel-cell cars every year. A related scheme is being used by the California state government to encourage sales of conventional electric vehicles. Governments could also help by offering drivers the chance to earn income-tax credits if they buy hydrogen vehicles. As use grows, the energy companies would have more incentive to expand the hydrogen distribution system.

Such a plan could get hydrogen-powered vehicles powered by fuel cells on the road, but

others argue that the process would be quicker if hydrogen were burned instead. All conventional engines powered by petrol, from turbines to cars, could be made to burn hydrogen with fairly minor alterations. Both Ford and BMW have developed vehicles that use hydrogen to power modified internal combustion engines (ICEs). Hydrogen ICEs still have to face the problem of developing an infrastructure for hydrogen distribution, but because only minor re-engineering of vehicles is needed, large-scale production would be cheaper and quicker than producing fuel-cell cars. Bob Natkin, leader of Ford's hydrogen ICE programme at the company's laboratories in Dearborn, Michigan, says he could have a hydrogen ICE available in three to five years. BMW is running on a similar timetable.

Work on modified ICE prototypes suggests that they will use hydrogen less efficiently than fuel cells. But their performance could be improved by using a 'hybrid' engine that also uses battery power. Petrol-run hybrid cars, which use batteries in conjunction with ICEs, already exist. Although they consume fossil fuels, their efficiency is around twice that of normal petrol vehicles. Hybrid vehicles running on

Cranking up the hydrogen flow



Going green: algae could be used to split water and so meet hydrogen demand.

organization that brings together academics and industry representatives interested in hydrogen technology. Reforming could be extended to any fossil fuel, or even plant matter, as an interim solution to the need for hydrogen.

But researchers are also working on methods for hydrogen production that avoid fossil fuels altogether. John Turner of the National Renewable Energy Laboratory in Golden, Colorado, leads one of several groups investigating photoelectrochemical systems. These combine photovoltaic materials, which use light to produce electricity, with electrolyzers, which split water. Rather than using a solar cell to generate electricity, which is then fed to a separate electrolyser, Turner's team aims to create single devices that can split water once submerged and illuminated.

Another group at the same lab is studying the green alga *Chlamydomonas reinhardtii*, which uses sunlight to split water and produce hydrogen and oxygen, in the hope of scaling up the process to create cultures of algae that can produce industrial levels of hydrogen.

Producing hydrogen is not difficult — passing electricity through water generates hydrogen at the negative electrode and oxygen at the positive electrode. But with existing renewable energy technologies not ready to deliver the electricity needed, other methods are needed.

'Reforming', or 'cracking', natural gas is currently the cheapest and most common method of hydrogen production. The technique is used by petrochemical companies to generate the hydrogen used in the synthesis of a variety of useful compounds. Using industrial-scale crackers, BP alone produces enough hydrogen to run over a million fuel-cell cars, says Jeff Rinker, the company's general manager for hydrogen and chair of the National Hydrogen Association, a Washington-based



Ray of hope: John Turner wants to use solar power to generate hydrogen.

W. GRETZ/NREL

R. PETERSON/NREL



Convertibles: hydrogen fuel tanks feed a converted internal combustion engine.

► hydrogen ICEs could achieve around 80% of the efficiency of fuel-cell vehicles, says Jay Keller, manager of the hydrogen programme at Sandia National Laboratories in Livermore, California. Modified ICEs do produce additional emissions of gases such as nitrogen oxides, which are pollutants and greenhouse gases, but Natkin says the emissions from his prototypes are already well below those from normal engines and could be reduced to close to zero.

Williams says that introducing the hybrid hydrogen ICE vehicles would offer the chance to work with a hydrogen infrastructure in preparation for the full-scale introduction of fuel cells. They would then be phased out over time.

These plans sound promising, but one major problem — that of storing hydrogen fuel — has to be overcome first. The Partnership for a New Generation of Vehicles, an initiative that brings together the US government and vehicle manufacturers with the aim of developing environmentally friendly vehicles, estimates that to satisfy consumer demand a car needs to run for around 600

kilometres on a single tank of fuel. Using current fuel-cell technology, around 5 kilograms of hydrogen would be needed to travel this distance. But this would require a pressurized tank of roughly 180 litres — far too large for family cars, which have tanks that hold an average of about 50 litres. Buses need even larger tanks but, unlike cars, have the space in which to put them.

Higher pressures can be used to cram the hydrogen into a smaller volume, but, as with any pressurized gas, this increases the risk of explosion, so numerous alternatives are under development. Storage density could, for example, be increased by adding metal hydrides to the tanks. These absorb hydrogen molecules, increasing hydrogen density without increasing pressure. The hydrides release the hydrogen when heated, allowing fuel flow to be controlled. Carbon nanostructures, such as nanotubes, could potentially be used in a similar way.

Tanked up

Researchers are currently trying to increase the amount of hydrogen these metal hydrides and nanotubes can hold. In many cases, commercial secrecy prevents the exact figures from being revealed. But Sigmund Gronich, who leads the US Department of Energy's Hydrogen Program Team, says that some government-funded groups have created mixtures of metal hydrides and hydrogen in which the hydrogen makes about 5% of the weight of the two. This would get the size of a fuel tank required for 600 km of motoring down to around 100 litres.

Although this is still too large for a family car, Keller argues that it could serve as a practical storage density if drivers can be convinced to accept the need to fill up more frequently. He suggests that a target of 320 km per tank, or around four hours' continuous driving, is reasonable. This would require a more practical 55-litre tank, at current hydride storage densities.

Meanwhile, some groups have claimed hydrogen-storage figures of 8% by weight for nanotubes, although others have found their claims hard to replicate (see *Nature* **410**, 734–735; 2001), and it is unclear how successfully these laboratory experiments would scale up into commercial fuel tanks. But with interest in both storage mechanisms currently intense, many in the field remain confident that the problem of hydrogen storage will be cracked.

With so many different approaches being pursued, some have argued that the money and time being ploughed into hydrogen research would be better spent on a more focused range of projects. Environmentalists add that oil and car companies need to invest more in the field. But overall, the drive for hydrogen power seems to be gaining momentum.

In Iceland, for instance, non-governmental groups and researchers, with the support of companies such as Ballard and Daimler-Chrysler, are working on a plan that would almost totally remove fossil fuels from the country's economy within 30 years. Iceland is in a unique situation. Vehicles will be powered by methanol — which can be derived from a by-product of the country's metal-production industry. Electricity needs can be met by abundant geothermal sources. The route to a hydrogen economy is not so straightforward elsewhere, but hydrogen supporters say that everyone will learn from Iceland's experience.

Charles Stone, a vice-president at Ballard, is in no doubt that other countries will follow Iceland's lead. "The goal of getting to a hydrogen-based system is so compelling, and the people involved in this are so ingenious, that I'm very confident we'll get there," he says.

Mark Schroppe is a freelance writer in Melbourne, Florida.

US Department of Energy Hydrogen Program

► www.eren.doe.gov/hydrogen/program.html

Partnership for a New Generation of Vehicles

► www.uscar.org/pngv



Desert manoeuvres: last year's Clean Energy World Tour, which demonstrated hydrogen-powered vehicles, started in Dubai.

The curtain falls

Theatre seemed the ideal way for US bioprospectors working in Mexico to tell local people about their work. But did the plays distract attention from the involvement of commercial interests? Rex Dalton reports.



Tough search: bioprospecting in the Chiapas region of Mexico has proved to be fraught with difficulty.

Three years ago, a team of US scientists set out for the lush highlands of southern Mexico. The biochemical secrets of the region's plant life, together with the local Maya people's knowledge of how to exploit them, promised a rich pharmaceutical bounty. By staging plays to explain their work, the researchers hoped to avoid the accusations of biopiracy that had dogged similar projects.

This October, after countless hours collecting specimens, the bioprospectors were forced to retreat home by the region's turbulent politics. As those involved lick their wounds, they and others are lamenting a lost opportunity.

The project was headed by Brent Berlin, an ethnobiologist at the University of Georgia in Athens who has decades of experience in Mexico. Together with researchers from El Colegio de Frontera Sur (ECOSUR), a university in the Chiapas region of Mexico, his team started to explore the local rainforest.

The research was part of the US International Cooperative Biodiversity Groups (ICBG) programme, which aims to identify plants with ingredients that could treat important diseases in the United States. As is usual for ICBG projects, Berlin's team had sketched out an agreement for distributing profits from the research. Any return was to be split equally between a trust for local people, ECOSUR, the University of Georgia, and MolecularNature, a company in Maidenhead, England, which set up a plant biology laboratory at ECOSUR and helped train local researchers.

But pressure groups such as the Rural Advancement Foundation International (RAFI), based in Winnipeg, Canada, attacked the researchers on the grounds that local people were not properly informed of

the project's aims. Initially, the researchers attempted to explain their aims by conducting interviews with local people, a strategy that had limited success. The local language — Maya — does not contain words for 'genes' or 'patents', and the researchers found it difficult to identify community leaders.

Dramatic moments

Theatre, which plays an important role in the local Maya culture, seemed to offer a solution. Around two years ago, Berlin's team, together with help from ECOSUR, moved between jungle hamlets, staging around 50 plays. Local people were recruited by ECOSUR researchers to act in the plays, which showed how plants were being collected and tested, and how villages could assist in the process and conserve the plants by starting their own botanical gardens. The actors also donned white coats to show how the researchers hoped to use the plants.

According to Robert Nash, research director at MolecularNature, who witnessed some of the performances, the plays appeared to work well in communicating their message to local people. Officials at the US National Institutes of Health (NIH), which helps fund ICBG projects, agree, saying the plays prompted most of the local communities to permit the bioprospecting.

RAFI, which this year changed its name to Action Group on Erosion, Technology and Concentration, together with local action groups, claimed that the plays were being used as a smokescreen to obscure the researchers' real objective — commercial exploitation.

Matters were further complicated when the new Mexican government, which was elected in July 2000, said that Mexico's bioprospecting legislation needed reviewing.

This review, which is still ongoing, undermined the legal status of Berlin's research and, together with RAFI's campaign, forced the NIH to ask Berlin to suspend collecting late last year.

Berlin responded by proposing to use the grant to develop standards for gaining informed consent from the Maya. NIH officials agreed, but criticism in Mexico continued to grow.

RAFI's attacks struck home with ECOSUR researchers, who became less willing to be associated with the research. They finally withdrew their support in October, effectively killing the project.

Berlin refuses to comment on recent events, but the project's end has saddened others involved. "It's a shame it had to go down," says Joshua Rosenthal, who directs the ICBG programme at the NIH's Fogarty International Center in Bethesda, Maryland. "This was an opportunity to test a model for ethical collaborations with indigenous people."

Regulated bioprospecting is seen by many as vital if the area's biodiversity is to be exploited in a sustainable manner. But companies such as Diversa, a San Diego firm that had invested in other Mexican projects and would have shared profits with its Mexican collaborators, have already turned their attention to other countries. And undocumented collecting by rogue companies, which scientists familiar with the area say has been going on for years, looks set to continue.

"You won't see any legal bioprospecting in Mexico for years," says Ignacio Chapela, a plant molecular biologist at the University of California, Berkeley. "This is a forgone opportunity for the Maya."

Rex Dalton is *Nature's* West Coast US correspondent.

D. LEHMAN/CORBIS; J. ROSENTHAL

Crops grown on set-aside land bring wild birds back to the fields

Monitoring is under way, and results so far are promising.

Sir — David Kleijn and colleagues report (*Nature* **413**, 723–725; 2001) that a Dutch scheme designed to improve agricultural grasslands for wading birds had failed — more birds were found on nearby conventionally managed grass fields — and call for scientific monitoring of the effects of similar schemes elsewhere.

Just this type of monitoring programme is being conducted in the United Kingdom, showing that the ‘wild bird cover option’ — the growing of certain crops on set-aside ground to benefit declining farmland bird populations — can indeed be successful.

The Game Conservancy Trust and the British Trust for Ornithology, supported by the Department for Environment, Food and Rural Affairs (DEFRA), are monitoring various crop mixtures sown on more than 100 farms across southern and central England, with the aim of determining which crops are preferred by the many species of birds encountered. Experimental work will confirm birds’

preferences and inform on the amount of cropping needed. Similar work is under way in Scotland.

So far, the findings show that all crop types studied support densities of farmland birds that are orders of magnitude higher than nearby conventional fields, with the biennial brassica kale (*Brassica napus*) and quinoa (*Chenopodium quinoa*), a member of the beet family — both seed-bearing plants — being the preferred crops for most birds. Given suggestions that the decline of farmland birds could be due to the recent loss of over-winter food supplies, the implementation of prescriptions such as ‘wild bird cover’ could have a major impact on bird populations in the United Kingdom and, no doubt, elsewhere.

Chris Stoate*, Dave Parish†

*Allerton Research and Education Trust, Loddington House, Loddington, Leicestershire LE17 9XE, UK

†The Game Conservancy Trust, 21 South Esk Road, Tannadice, Angus DD8 3SH, Scotland

Schemes are monitored and effective in the UK

Sir — Kleijn *et al.* call for reliable evaluation of agri-environment schemes across Europe¹, after concluding that schemes in the Netherlands did not increase the numbers of target wildlife species.

Detailed monitoring of this type has been undertaken in the United Kingdom since 1987. The results of various surveys have been published in books (see, for example, ref. 2) and journals (for example, ref. 3) and — for England — on the Department for Environment, Food and Rural Affairs’ website (www.defra.gov.uk/erdp/erdpfrm.htm).

These reports and papers demonstrate that, in general, the UK schemes have maintained biodiversity, and show some limited enhancement in wildlife. When monitoring has revealed problems, management prescriptions given to farmers have been changed to improve the effectiveness of the scheme concerned.

Although the UK government and its agencies have spent considerable sums of money to determine the efficacy of agricultural–environment schemes in the United Kingdom, governments are often reluctant to fund adequate monitoring

because of the administrative costs — the more money that is spent on monitoring, the less there is to advise farmers.

Unfortunately, the costs of administering European agricultural–environment schemes have to be borne by the country concerned, and are not subsidized by the European Union (EU). If the EU is willing to pay for monitoring, it will help to ensure the adequate assessment, and indeed the comparison, of such schemes.

Peter D. Carey

Centre for Ecology and Hydrology, Monks Wood, Abbots Ripton, Huntingdon PE28 2LS, UK

1. Kleijn, D., Berendse, F., Smit, R. & Giliissen, N. *Nature* **413**, 723–725 (2001).
2. Sheldrick, R. D. ed. *Grassland Management in Environmentally Sensitive Areas* (British Grassland Society, Lancaster, 1997).
3. Peach, W. J. L., Lovett, L. J., Wotton, S. R., Jeffs, C. *Biol. Conserv.* **101**, 361–373 (2001).

War: no time for dissent

Sir — Your correspondent Morton K. Brussel is wrong (“Don’t use terrorism as an excuse for militarism”; *Nature* **414**, 249; 2001). We are not in a new war promoted by the Bush administration. We are responding to an unprovoked attack that is a threat to civilized life. This is understood by the whole world, and we have the support of all the leading powers, east and

west. To call this “yielding to the politics of an uncivil society” is amazing nihilism. I do not understand such ignorance in our academic community.

Arno Arrak

5 Chatham Place, Dix Hills, New York 11746, USA

Church backing depends on ethical use of animals

Sir — As reported in your News story “Vatican approves use of animal transplants ‘to benefit humans’” (*Nature* **413**, 445; 2001), the Vatican’s Pontifical Academy for Life has recently organized several meetings to discuss the science, theology, ethics, law and procedures for xenotransplantation (see www.academiavita.org for a complete account and references). Here, we draw attention to the Vatican’s conclusions, and offer a Catholic view on two current debates.

The Vatican sees three issues as critical from a theological perspective. First, man has a right and duty to act within and on the created order, making use of other creatures to achieve the final goal of all creation: the glory of God through the promotion of man.

Second, in xenotransplantation the service of animals to man represents a totally new application that is not in conflict with the order of creation. Humans must answer to the Creator for the manner in which they treat animals. The sacrifice of animals is justified only if it is required to achieve an important benefit for man, including experiments on animals and/or genetic modification of them.

Third is the question of whether an animal organ in the human body modifies a person’s identity. Although transplantation of the encephalon or the gonads, which are integral to personal identity, can never be morally legitimate, that of purely functional organs is legitimate.

Finally, we mention two ethical issues. First, the vast healthcare resources that would be used for xenotransplantation are justified by the urgent need to save the lives of patients. Second is the ethical need to acquire correct information on potential benefits and risks. This should be communicated to the public. By public discussion and debate, society should help to identify the conditions under which it is acceptable to invest resources and hope in this new therapeutic approach, in the light of scientific uncertainties and the urgent need for more organs for transplantation. **Archbishop Monsignor Elio Sgreccia*, Don Maurizio Calipari*, Marialuisa Lavitrano†**

*Pontifical Academy for Life, Via della Conciliazione 3, 00120 Vatican City, Italy

†Department of Experimental Medicine, University of Milano-Bicocca, Via Cadore 48, 20052 Milan, Italy

Out of thin air

Towards an understanding of our responses to high altitude.



High Altitude: An Exploration of Human Adaptation

edited by Thomas F. Hornbein

& Robert B. Schoene

Dekker: 2001. 982 pp. \$235

Sue Jackson

Hippocrates, Plato and Aristotle knew nothing of haemoglobin, oxygen or carbon dioxide, let alone aerobic metabolism, yet in essence their conviction that air circulated though the heart was not wrong. Respiratory physiology is necessarily integrative, and the effects of existing at high altitude are sensed and responded to at every level — by cells through whole bodies to populations. So this volume quite appropriately encompasses the function of almost every organ system except the skeleton and gut, and is unified by focusing on hypoxia — “a surfeit of too little oxygen” — rather than lung function.

We are drawn to worship, live among and climb great mountains, and in so doing we have become fascinated by our bodies' responses to them. This has led to a huge and rapidly expanding literature through which the multiple authors of this volume pick their separate ways, referring to one another's chapters with varying degrees of specificity and accuracy. *High Altitude: An Exploration of Human Adaptation* is just that — a peregrination with a certain geographical bias. Of 45 contributors, 33 are from the United States, with three each from Canada and South America, four from Europe and one each from the United Kingdom and Australia. Does this matter? Tales of exploration are consistently

chauvinist but no less enjoyable for that.

The higher one goes above sea level, the fewer oxygen molecules per unit volume there are in the air one breathes. The human hyperventilatory response to such “subtle and delicate aire”, as the sixteenth-century Spanish Jesuit priest Jose de Acosta described it, flushes carbon dioxide out of the body, leading to reduced hydrogen-ion concentration — respiratory alkalosis. The body fluid balance and the kidneys are affected. Blood capillaries in the lungs constrict and their walls thicken, elevating pulmonary blood pressure.

If breathing cannot supply enough oxygen to the tissues, the body tries to compensate by synthesizing more red blood cells. But the blood may then eventually become so thick that it fails to circulate properly, causing tissue hypoxia and, paradoxically, stimulating the synthesis of yet more red blood cells. This syndrome (called chronic mountain sickness or Monge's disease after Carlos Monge, the author of one of the chapters) highlights the limits of human function at altitude.

Other potentially fatal consequences of life at high altitude — among them peripheral, pulmonary and cerebral oedema, reduced birth weight in babies, and the complication of pregnancy known as pre-eclampsia — all attest to human maladaptation. Although we live and sojourn at altitudes higher than 4,000 metres, our homeostatic mechanisms evolved closer to sea level and often fail at altitude. When they do not, cellular oxygen-sensing pathways lead to physiological changes that result in

lifelong acclimatization and, if survival and reproductive success are affected, to heritable changes.

Humans are recalcitrant subjects for the evolutionary physiologist investigating hypoxia adaptation to deal with — they ask questions, and (hopefully) have the freedom to withdraw from experiments. Individuals and whole populations migrate across and between continents, from lowlands to highlands and back again to interbreed. Environmental factors such as diet (and hence iron status and metabolism) and level of activity influence many of the physiological systems involved in the respiratory cascade from our lungs to our energy-producing mitochondria. But the fascination prevails, and several of the book's chapters supply fine discussions of the compounding effects of culture, food intake, metabolism and exercise on responses to hypoxia.

The volume seems fragmented, perhaps reflecting the field itself. Although Monge highlights the need for more research on hypoxia's effects on the morphology and density of blood capillaries in tissue, he does not cite the extensive work of Hans Hoppeler, Bengt Kayser and colleagues on this subject. A discussion by Curtis Smith, Jerome Dempsey and Thomas Hornbein of the way in which our breathing responds to hypoxia seems to me incomplete with only one reference to the extensive work by Peter Robbins and his colleagues in Oxford. Are such omissions an inevitable consequence of the size of this literature and particular axes of academic communication? I prefer to think that they need not be so.

KEREN SU/CORBIS

For readers already familiar with the literature on high-altitude physiology, this volume will summarize a set of authoritative perspectives based on more than five decades of intensifying multidisciplinary research. For the interested physiologist or human biologist, it is as extensive an entry point into the literature as any yet published. It shows that modern physiologists now face uncharted territory as vast as that which confronted the Danish physiologist Christian Bohr when he described the oxygen–haemoglobin dissociation curve in 1910.

However, a few new tools are to hand. With an expanding array of non-invasive imaging techniques, we can examine substrate turnover in living brain and muscle tissue. With caution, we can reach inside the cells of model organisms to examine pathways of responses to hypoxia from the outer plasma membrane to messenger RNA and protein synthesis. We can induce people to cycle in hypobaric chambers that artificially lower barometric pressure (and hence oxygen availability) to levels equivalent to those at the summits of the world's highest mountains, and then biopsy their muscles. We can use familial studies and DNA analyses to trace the heritability of physiological traits and how long ago populations diverged, factors that help us to trace patterns of genotypic adaptation to altitude.

On old maps, unexplored ground was sometimes marked with the words "Here be dragons". In their preface the editors express a wish that this volume should reveal gaps in current knowledge. Although politically troubled eastern Africa is home to the high-altitude peoples of this continent, we know nothing about their physiological responses to hypoxia. Comparisons between Andean and Himalayan peoples, together with the little we do know about Kenyans, suggest that the addition of African data to this picture would be very worthwhile. And why does the affinity of the human haemoglobin molecule for oxygen seem so much less variable than that of mice — or has its variability just not been adequately explored? Much has been revealed about cellular sensing of hypoxia in the past five years, but the exact nature of the sensing pathway is still elusive. Are the key metabolic downregulators in the hypoxia response similar to those for other states of reduced metabolic demand? How do genes, physiology and culture interact to produce people who climb Mount Everest without supplementary oxygen and athletes from remote Rift Valley villages who come to dominate the world athletics circuit? These are the dragons I spotted. Because every reader will see others, Hornbein and Schoene have accomplished what they set out to do. ■

Sue Jackson is in the Department of Physiological Sciences, University of Stellenbosch, Private Bag XI, Matieland 7602, South Africa.

Lifting the curtain on the Nobels

The Politics of Excellence: Behind the Nobel Prize in Science

by Robert Marc Friedman

Times Books: 2001. 379 pp. \$30

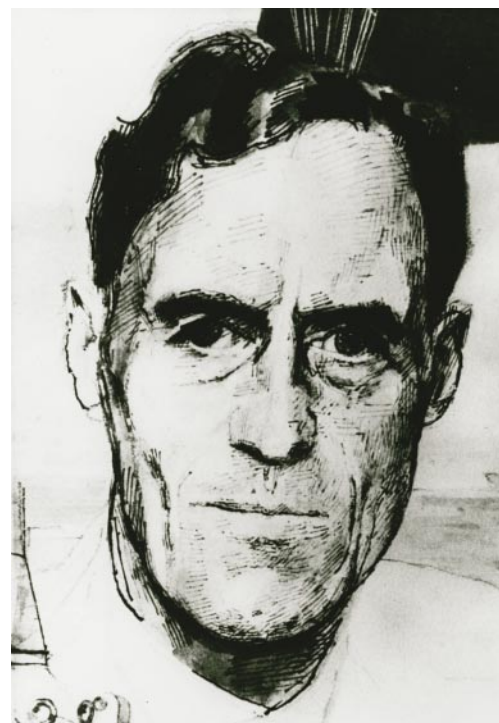
Aant Elzinga

Alfred Nobel, the wealthy inventor of dynamite, died on 10 December 1896. Five years later a practice began that will acquire added lustre in this year of its centennial: Nobel's last will instructed that his fortune be invested in safe securities, with the interest used annually for five prizes, to "those who, during the preceding year shall have conferred the greatest benefit on mankind". One prize was for "the most important discovery or invention within physics", another for the "most important chemical discovery or improvement".

There was a problem, though. It is impossible to recognize immediately what will become the most important pioneering contributions to science. Time is needed for the dust to settle after a "discovery" or an "invention". On the other hand, there is a danger in letting too much time pass before making an award. Thus, the Nobel Foundation had to develop 'bylaws', later updated and modified, to define limits and allow greater precision in interpreting the testament. Robert Friedman, a historian of science at the University of Oslo — after many years of spadework in the Nobel archives — presents us with a book that is far more than just a history of the first half-century's prizes.

Friedman is interested in the origins of the prize, its role in Swedish culture and identity, and in the politics behind the awards. So he goes behind the scenes. How did the executors come to grips with Nobel's testament, interpret it and set up the rules of the game? How have the personalities involved influenced the procedures? The division of labour between members of the physics and chemistry committees and the Swedish Royal Academy of Sciences — early on and later — was not always clear cut. Friedman traces the negotiations at every level, including not only the bitter conflicts over particular candidates for the prizes, but also the ramifications of geopolitical turbulence and pressures during and immediately after the First World War, and in the wake of the Second World War.

Conflict between scientific traditions is another theme. From the outset, Uppsala University jealously guarded an older, experimentalist ideal of science, narrowly interpreted, whereas leading professors at Stockholm pushed for a liberalization of the interpretation of notions such as "invention" and "discovery" to make room



Postwar winner: Patrick Blackett, particle physicist and war hero, received the 1948 award.

for theoretical work. The case of Albert Einstein, long refused a prize, and then at last receiving his prize for the discovery of the law of the photoelectric effect — not relativity theory — reflects these tensions. Another controversy surrounded the chemistry award to Fritz Haber, because of his role in German gas warfare. This revealed the significance of further factors: one was the pro-German sentiment in Sweden, and another, concern about reconciliation and rebuilding internationalism after the rupture of scientific communities caused by the First World War.

The 1920s brought new quarrels over changes to the way the prize was managed and awarded. Astrophysics was declared out of bounds, and some clever lobbying, led by the theoretician Carl Oseen, brought the 1925 physics prize to Sweden's Manne Siegbahn. This was part of a hidden domestic agenda by a section of the physics community, particularly Oseen, to build up Swedish atomic physics.

Resistance to the new quantum theory of Erwin Schrödinger and Werner Heisenberg, and their breakthrough onto the Nobel prize rostrum, also makes fascinating reading, as do the stories of the campaigns and counter-campaigns that followed, in both physics and chemistry. The award of the postwar physics prize to England's Patrick Blackett, a war hero and left-wing critic of postwar militaristic nationalism and nuclear weapons, is said to have been an overt case of politics by other means.

Apart from international and national politics, Friedman argues, the complex

micropolitics of defining terms of reference for prizes are constantly in play — who writes the review of potential candidates, and the jockeying in the committees and in the academy's plenary meeting that influences the outcome of the process. The academy has been known to override committee recommendations, and infighting, ignorance and fixation on pet ideas were perhaps more prevalent in the early days, when review procedures were less systematic and a few dominant personalities held sway. But Friedman shows how, even today, when one looks behind the scenes the Nobel prize belies the popular myth of being a dispassionate, neutral, disinterested process.

Friedman's tale is never dull. Although not intended as a scholarly work, the book is not mere gossip. Aimed mainly at a broad audience, it nevertheless has a separate section of elaborate notes for historians and students of the politics of science wishing to delve deeper. What irritates is the cuteness of some subheadings put in to engage the popular reader. Also, with all the focus on conflicts and controversial events, one wonders what happens in the cases of more peaceful convergence around prizewinning candidates, or whether, in such an extraordinary setting, those don't exist. ■

Aant Elzinga is in the Department of the History of Ideas and Theory of Science, University of Göteborg, SE 405 30 Göteborg, Sweden.

Sex appeal of a musical insect

Katyids and Bush-Crickets: Reproductive Behavior and Evolution of the Tettigoniidae
by Darryl T. Gwynne

Cornell University Press: 2001. 344 pp.
\$42.50

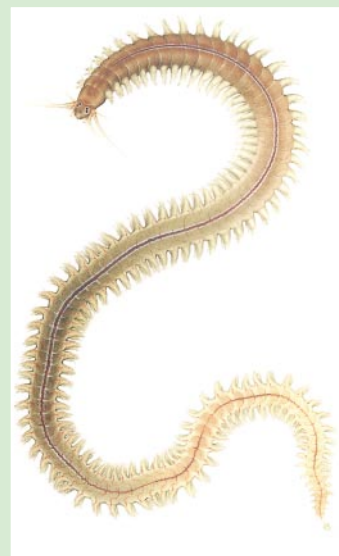
Marlene Zuk

With great reluctance, I have decided against my initial impulse to give this book to everyone I know for Christmas. Some friends might appreciate it, but I can't quite see my Texas in-laws expressing much enthusiasm. This caveat notwithstanding, *Katyids and Bush-Crickets* is a fascinating and enjoyable read, full of both charming trivia and scholarly insight. The book is brimming with 'did-you-knows?': for example, the term diapause (the temporary cessation of development in insects) was coined by the renowned Harvard expert on ants, William Morton Wheeler, in his 1893 thesis on katyids. The book is also beautifully illustrated, with many line drawings, breathtaking colour plates, and figures that are all redrawn from their original source — a welcome touch of clarity and uniformity that is absent

A turn-up for worms

Here are two examples of the "multiple variations on a theme" that comprise the diverse species of marine worms, or polychaetes — *Eupolymnia* (right) and *Hediste diversicolor*.

Polychaetes by Greg W. Rouse and Fredrik Pleijel (Oxford University Press, £99.50, \$175) examines these animals, which are found in nearly every marine habitat.



in many such books. Virtually all of the chapters begin with pertinent and amusing quotations from sources ranging from Oliver Wendell Holmes to Dean Martin.

Why is this group of insects — called katyids in North America and bush-crickets in Europe — so interesting? Although Darryl Gwynne provides information on diversity, evolution and the economic importance of the family, the focus of the book, and the answer to the question, is the katyid sex life. Unlike most insects, and indeed most animals, katyid males contribute more than just sperm when mating. They generally provide females with a nutritious parcel, the spermatophylax, which is attached to the sperm-containing ampulla. The female consumes the spermatophylax after mating, while the sperm move into her reproductive tract.

This nuptial gift can represent up to 40% of the male's body mass, and hence is not easy to produce. Males may mate only once every few days, garnering the resources to manufacture one of these massive meals between mating. The hefty donation has important implications for sexual selection, because the general rule is that the reproductive success of males is limited only by the number of females they can inseminate, rather than by the investment they make in producing offspring. This is likely to be far more variable

than the factor that generally limits female reproductive success — the number of offspring they can produce and (where appropriate) rear. Because females in a species usually make the greater investment, sexual-selection theory predicts that females will benefit by maximizing the quality of their offspring, but males will benefit more from maximizing the quantity.

All that is reversed in katyids and in several other organisms, including sea-horses and their relatives the pipe-fish, as well as a handful of birds and insects. If males do not merely produce a dab of semen, but must proffer a costly gift to females before mating, they would be expected not only to be choosy about their mate, but also to be sensitive to the availability of resources in the environment in a way that males in other species are not. Such a reversal of sex roles provides a crucial test of current theory on how the sexes are expected to differ.

How did the spermatophylax evolve? Gwynne evaluates two hypotheses. The paternal-investment hypothesis suggests that males provide females with a meal because it increases the fecundity of the female or the quality of the offspring she produces, and hence increases the male's fitness along with that of the female. According to the ejaculate-protection hypothesis, in contrast, the spermatophylax serves to keep

females busy eating it while sperm are transferred. Males providing larger gifts therefore gain a larger proportion of fertilizations.

In a careful dissection of experiments on a variety of species, Gwynne concludes that the paternal-investment hypothesis has received the more support; in many katydids, the ampulla is completely empty long before a female has finished her gelatinous meal.

Nevertheless, much remains to be discovered about katydids. Indeed, one of the book's strengths is the prevalence of ideas for further research — there's a possible dissertation project on virtually every page. Why, for example, have so few katydid species lost acoustical function, compared with their equally musical relatives the crickets? What favoured the occasional evolution of parthenogenesis in this unlikely-seeming group? Why do some katydids mate only once? The section on risks and survivorship is particularly dramatic: "in nature, death usually comes from a plethora of hostile forces", including parasitoid flies that home in on the male's calling song, horsehair worms that burst from the body cavity like lethal spaghetti, as well as bats, frogs, birds and fungi.

One can quibble with any work of this magnitude, and I did my share. The references on a few topics, such as the role of sexual selection in speciation, were curiously outdated. Sexual-conflict theory — the conflict between males and females over mating — is a relatively new, hot topic in sexual selection, but is mentioned rather briefly; its explicit incorporation into ideas about spermatophylax evolution might have been fruitful. Gwynne draws a distinction between so-called passive and active female choice, a dichotomy I have always disliked. These bits of carping aside, *Katydid and Bush-Crickets* is a book to share with all students and professionals in ecology and evolutionary biology — and, who knows, maybe also with a few relatives at Christmas. ■

Marlene Zuk is in the Department of Biology, University of California, Riverside, California 92521, USA.

More biology

Carnivore Conservation

edited by John L. Gittleman, Stephan M. Funk, David Macdonald & Robert K. Wayne
Cambridge University Press, £34.95, \$49.95 (pbk)

Avian Ecology and Conservation in an Urbanizing World

edited by John M. Marzluff, Reed Bowman & Roarke Donnelly
Kluwer, £112, \$159.95

Fungal Conservation: Issues and Solutions

edited by D. Moore, M. M. Nauta, S. E. Evans & M. Rotheroe
Cambridge University Press, £65, \$95

Science in culture

Scope for sentiment

Joseph Wolf's illustrations of animals in action.

Martin Kemp

There seems to be a sharp contrast between the Victorian fondness for sentimental depictions of anthropomorphized animals and the hard objectivity of professional natural history in the mid-nineteenth century. Little apparently links the tough impersonality of darwinian 'natural selection' and the personalized beasts of Sir Edwin Landseer, famed for such paintings as *The Monarch of the Glen*. When we find these two poles represented in the work of one artist-illustrator, Joseph Wolf, we may wonder how he was able to accommodate such visual schizophrenia. I believe his work actually points us towards a reassessment of both the sentiment and the science.

Born and educated in Germany, Wolf enjoyed an immensely successful career in London from 1848 to 1899. He collaborated with some of the great naturalists of his day, including John Gould, Philip Gosse, Herman Schlegel and Daniel Elliot, providing them with great coloured lithographs of striking precision, and he long served as the chief illustrator for the *Proceedings of the Royal Zoological Society*. Yet his paintings, exhibited at the Royal Academy in London, tell affectingly emotional stories of the struggles between predator and prey in often-hostile environments, and his popular-book illustrations verge on animal capers.

The peak of Wolf's popularity came with *The Life and Habits of Wild Animals*, published in 1873 as a grand picture book with 20 woodcuts. *Inquisitive Neighbours* is typical of both the narrative quality of the illustrations and their humanizing titles. When we realize that the text — no less evocative and vivid in bringing the characters of the animals to life — was by the eminent American naturalist Daniel Elliot, for whom Wolf had previously undertaken five suites of illustrations, the schism between popular art and observational science is less clear cut.

Wolf himself claimed that, to depict creatures in lifelike action and interaction, the artist needed an accurate knowledge of the structures that give rise to their motion. He was scathing of those who know only "the map of an animal", such as those "ornithologists who don't recognize nature — don't know a bird when flying. A specimen must be well dried before they recognise it."

Even the apparently 'arty' expressions join seamlessly with scientific concerns, in particular with Wolf's engagement as one of the illustrators for Darwin's *The Expression of the Emotions in Man and Animals*, published just a year before his own *Wild Animals*. Darwin provides unexpected succour for those who sensed a similarity between the way human and animal emotions are manifested:

"No doubt as long as man and all other

animals are viewed as independent creations, an effective stop is put to our natural desire to investigate as far as possible the causes of expressions. With mankind some expressions, such as the bristling of the hair under the influence of extreme terror, or the uncovering of the teeth under that of furious rage, can hardly be understood except on the belief that man once existed in a much lower animal-like condition."

When we react with empathy to the threatened dove's puffed out chest and the alertly raised 'eyebrows' of the intruding squirrels, we are playing precisely to the kind of basic expressions for which Darwin was defining the mechanisms.

If the more sentimentalizing aspects of Wolf's art can be shown to adhere to scientific research, so we might surmise that Darwin's own picture of a nature in full-blooded interaction is unthinkable without the Romantic vision of those earlier artists, such as George Stubbs and Eugène Delacroix, who precociously delighted in vicious competition between the powerful and the weak, and the armed and the defenceless. ■
Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.

Visualizations: The Nature Book of Art and Science is a collection of essays edited by Martin Kemp (published by Oxford University Press and the University of California Press; £20, \$35).

Wolf's Inquisitive Neighbours, typical of his evocative depiction of animals.



mRNA readout at 40

John F. Atkins and
Raymond F. Gesteland

On 30 December 1961, 40 years ago, Francis Crick and colleagues published the results of a remarkable experiment that established without doubt the general features of the genetic code for protein synthesis. It was already known that the amino-acid sequence along the polypeptide chain of a protein is determined by the bases along the nucleic acid (DNA or RNA) of the genetic material, and that the code for this correspondence is non-overlapping along the base chain. Crick *et al.* established that a group of three bases encodes one amino acid, and that the chain of bases is 'read' in a sequential manner without punctuation from a fixed starting point.

These findings, in conjunction with the non-overlap, mean that a 'reading frame', set by the starting point, determines which particular sets of three bases specify the different amino acids (see figure) — a brilliant culmination of many years' work originating different coding schemes and trying to distinguish between them. In today's era of cloning, rapid sequencing and atomic-resolution structures, it is salutary that this cornerstone was put in place not by ordering a kit or sending samples to a sequencing centre, but rather through simple genetic crosses of a bacteriophage, combined with great insight in interpreting

the properties of the resulting mutants.

The extent to which certain phage mutants could be reverted to wild type by different mutagens revealed a class of 'deletion' (minus) and 'insertion' (plus) mutants. Crosses between representatives of the two classes (plus/minus or minus/plus) would give wild-type progeny because the reading frame must have been restored, whereas crosses within a class (plus/plus or minus/minus) gave no wild-type progeny — the reading frame remained disrupted (showing, incidentally, that the code was not based on two or a multiple of two bases). The crucial insight came from the discovery that a triple mutant within either class (minus/minus/minus or plus/plus/plus) could give wild type: the frame had been restored. Hence the reading unit, or codon, must be three (or a multiple of three) bases.

In the 40 years since this dramatic discovery, the detailed mechanism of the decoding of the non-overlapping sequential triplets is still not understood, despite recent success in obtaining atomic-level structural information about the subunits of the ribosome. The ribosome is a gigantic complex of more than 50 proteins and RNA, the latter having the primal role in the decoding process. Understanding the dynamics between the ribosomal RNA, transfer RNA and messenger RNA (mRNA) that establish and maintain reading frames in the ribosome remains a challenge today.

Evolutionary variants of triplet readout that exploit alterations of reading frames, fascinating in their own right, can also provide clues to how the standard process works so reliably. In the decoding of a minority of genes, signals in the mRNA dictate that a proportion of translating ribosomes shift, at a specific site, into one or other of the alternative two reading frames (see figure). The resulting trans-frame protein product is often produced in a set ratio to that synthesized by standard decoding. One example of such a product is the GagPol precursor of HIV and many other retroviruses.

In other cases, the programmed frameshifting occurs early in the readout process, providing a regulatory mechanism in which the efficiency of the shift responds to cell physiology; that is, it acts as a 'sensor'. Antizymes, negative regulators of polyamine levels, are an example of this process.

Nearly all the frameshifting used for gene expression involves moving the reading frame one base either 'backwards' or 'forwards' on mRNA. In rare cases, a constellation of signals causes decoding to resume after a block of nucleotides are traversed without specifying any amino acid. Signals in mRNA specify the resumption site. The current record distance

Genetic code

The importance of the concept of non-overlapping triplet reading of the genetic code, proved 40 years ago, remains huge.

efficiently 'bypassed' this way is a 50-nucleotide non-coding sequence. There is even a case in all bacteria where the resumption of decoding after interruption is on a different, specialized RNA.

These gymnastic feats are not the only way in which standard genetic readout is locally reprogrammed. Signals in mRNA can redefine 'stop' codons to specify an amino acid. Sometimes, this is to permit ribosomes to access coding sequence from further on, in which case the identity of the amino acid specified by the redefinition is unimportant. In other cases, redefinition is used to specify selenocysteine, the twenty-first directly encoded amino acid.

Apart from these cases of code extension where standard readout is dynamically and transitorily overridden (recoding), the standard reading frame is maintained with high accuracy. Errors in frame maintenance have a high cost, as they result in truncated proteins. Truncated proteins also derive from substitution mutants that create a premature stop codon: this type of mutant is responsible for about a quarter of human genetic diseases. Nevertheless, errors that give read-through of a premature stop codon, and certain framing errors near a frameshift mutation can lead to a small amount of full-length protein. The low efficiency of read-through of a stop codon can be increased by aminoglycoside antibiotics that decrease the fidelity of protein synthesis. The potential of this process for ameliorating genetic disease is now being assessed in clinical trials.

The importance of the concept of the principles of genetic decoding, proved 40 years ago, remains huge. Even when current advances in molecular biological and structural tools reveal the dynamics of its mechanism in ribosomes, the permutations will continue to challenge and surprise us. ■

John F. Atkins and Raymond F. Gesteland are in the Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112-5330, USA.

FURTHER READING

Crick, F. H. C. *et al.* General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
Crick, F. *What Mad Pursuit. A Personal View of Scientific Discovery* Ch. 12, Triplets (Basic Books, New York, 1988).
Judson, H. F. *The Eighth Day of Creation* (Simon & Schuster, New York, 1979).
Atkins, J. F. *et al.* Dynamics of the genetic code. In *The RNA World* 2nd edn (eds Gesteland, R. F. *et al.*) 637–673 (Cold Spring Harbor Laboratory Press, New York, 1999).

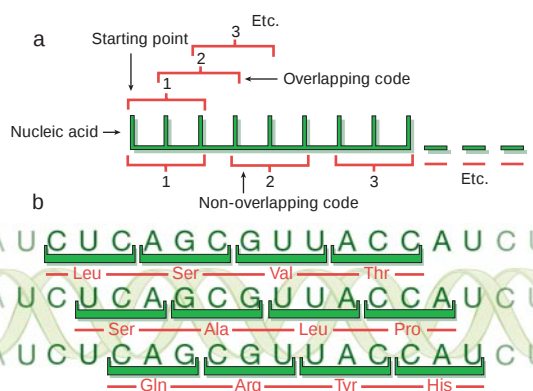


Figure 1 Reading frames. a, The difference between an overlapping code and a non-overlapping code. The short vertical lines represent the bases of the nucleic acid. The case illustrated is for a triplet code. b, Three possible reading frames in protein synthesis. A sequence of nucleotides in RNA is read in sequential sets of three nucleotides and is thereby translated into amino acids. The same RNA sequence can therefore specify three completely different amino-acid sequences, depending on the 'reading frame'.

Waving goodbye to measles

Peter M. Strebel and Stephen L. Cochi

An innovative way of analysing statistics on measles incidence in England and Wales since 1944 reveals recurring waves of infection originating in large cities. The information can guide strategies for preventing the disease.

Measles tends to occur in epidemics. It causes severe complications, such as pneumonia and blindness, and often kills. And it is highly communicable: in a population of susceptible individuals, a single primary case of the disease can result in 15–18 secondary cases¹.

The usual way of depicting the course of measles as it moves through a population is the epidemic curve—a graph of the number of cases occurring in a given population over time. Instead, as they describe on page 716 of this issue², Grenfell *et al.* have applied advanced methods of time-series analysis to weekly records of measles cases from England and Wales for the period 1944–1994. They show that recurring waves of measles infection begin in large cities and travel to peripheral towns. This fresh view of measles transmission in time and space provides the best documentation yet of what had previously been known largely anecdotally. It has implications for how we try to prevent, and eventually eradicate, the disease.

Uptake of measles vaccine in a population is the main factor affecting transmission of the disease; as uptake increases, epidemics become smaller and less frequent. In the 1990s, the Pan American Health Organization (PAHO) pioneered a highly effective approach to measles control. The approach combines high vaccination coverage for each group of newborns with periodic nationwide vaccination campaigns that target older children who either missed their first dose of vaccine or failed to develop an immune response³. When properly implemented, this strategy virtually halts measles transmission, as seen in various countries where it has been tried^{4,5}, and as is dramatically depicted in Fig. 1 for Romania. The few cases that are confirmed after such nationwide vaccination campaigns can be explained largely, if not exclusively, by import of the disease from countries where it remains prevalent.

In England and Wales, vaccination was initiated in 1968 and a nationwide, school-based campaign in 1994 has resulted in the sustained interruption of measles transmission since then⁶. Using data from the pre-vaccine period, Grenfell and colleagues² identify the cities of London, Manchester and Liverpool as the source of periodic waves of measles infection and the site of persistent reservoirs of the disease between epidemics. In the years 1945–1947 and 1962–1965 the

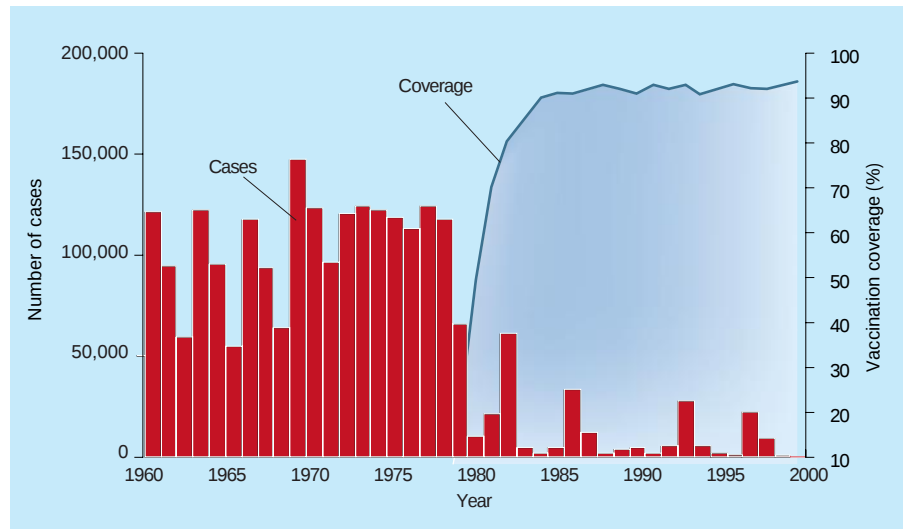


Figure 1 Beating measles with vaccination, as shown here for Romania. The red bars indicate the number of measles cases, and the blue line indicates vaccination coverage, among one-year-olds. In 1979, a single-dose measles vaccination strategy was introduced; a second dose was added in 1994; and in 1998 a nationwide mass-vaccination campaign was conducted among everyone between the ages of 7 and 18. In 1998, 1999 and 2000 the numbers of reported measles cases were, respectively, 9,547, 240 and 35. (Source: Ministry of Health, Romania, and ref. 10.)

inter-epidemic period was reduced to less than two years in London and the surrounding areas: this, the authors suggest, was due to increased birth rates (the 'baby boom' years) that fed susceptible people into the population at a higher rate.

Herein lies the main message of the paper. It supports the idea that preventing measles

in the world's most rapidly growing megacities, which are characterized by high birth and migration rates, is the greatest challenge to eradicating measles worldwide⁷. Although the PAHO strategy has been effective in stopping measles transmission in places such as Mexico City, São Paulo and Buenos Aires, none of these population centres will grow

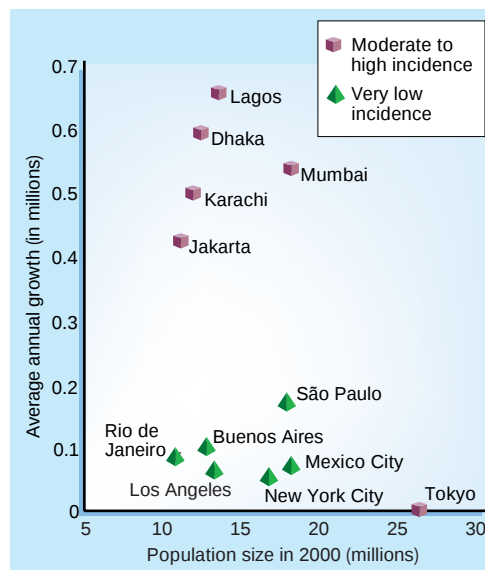


Figure 2 Measles and megacities. In 2000, the cities shown by green triangles had a low incidence of measles because of measles elimination strategies; these cities also have relatively low population growth. The other cities (purple squares) all had a moderate to high incidence of measles in 2000; in all of them, only the standard one-dose vaccination strategy was used, and all, except Tokyo, have an average annual population increase of at least 200,000 people. (Source: ref. 11.)

as rapidly as other cities including Mumbai (India), Dhaka (Bangladesh) and Lagos (Nigeria) (Fig. 2). Full implementation of the PAHO strategy is needed in the megacities and the surrounding regions to stop measles transmission. If this proves to be impossible, new approaches such as development of new vaccines for administration in early infancy may be needed.

Grenfell and colleagues² show that, in the pre-vaccine era, waves of measles infection moved away from London at a speed of around 5 km per week and were evident up to 30 km away from the city. There was a time lag between the peak in the originating urban centre and its arrival in the periphery — the lag was greater for smaller and more distant towns. Given this pattern, it is tempting to think that the extent of measles outbreaks can be controlled by vaccination ahead of the epidemic. But attempts to control outbreaks show that once a wave of infection has been established, it is difficult to stop it through selective vaccination. Often the intervention comes too late or, if it is in time, the 'wave' of measles infection still sweeps over and around the 'breakwater'. Current recommendations are that measles outbreaks should be prevented rather than controlled after they have started⁸. But vaccination in areas next to districts in which an outbreak occurs may be effective if high coverage of all susceptible age groups is achieved.

From Grenfell and colleagues' results, it might seem that concentrating on cities would be an effective vaccination strategy (although this is not something they themselves propose). This approach — based on the idea that cities are the key disease reservoirs — has been tried in Africa, and has not worked. In the late 1990s, Burkina Faso and Mozambique conducted mass vaccination campaigns, targeting children of preschool age in the largest cities with the aim of preventing spread to more remote rural towns and villages that health services found hard to reach. In both countries, the urban campaigns were associated with some reduction in the size of the expected epidemic in the city but were unsuccessful in preventing disease spread to rural areas⁹. The urban campaigns still left enough susceptible city dwellers to result in an epidemic, and then human migration seeded the periphery, leading to large outbreaks elsewhere.

These examples reinforce the need for universal vaccination against measles for all children regardless of where they live. The sheer number of people moving around for business or pleasure, in addition to forced migrations, make importation of measles a common occurrence: achieving and maintaining high population immunity by vaccination throughout an entire country is the only proven way of preventing large outbreaks.

The results in Grenfell *et al.*'s paper should stimulate activity in at least two areas. First, it is likely that the city-to-village dynamics that they document for measles also apply to other common human diseases. Among them are some that can be prevented by vaccines (rubella, mumps, varicella and hepatitis A) and some that, as yet, can't (those caused by respiratory syncytial virus, rotaviruses and adenoviruses), and influenza, which has pandemic potential but is vaccine-preventable. Better characterization of the patterns of disease transmission in space and time could offer unique opportunities for prevention: in particular, more research is needed on the effect of long-range jumps of infection that occur, for example, with air travel.

Second, the new work highlights the value of careful collection, analysis, dissemination and storage of disease-surveillance information. Accurate records of measles cases were kept in England and Wales long before vaccination against the disease was introduced, and Grenfell and colleagues' sophisticated time-series analyses were possible only because of the geographically detailed and complete nature of the data sets. We would be wise to strengthen the disease-

surveillance systems in both developed and developing countries, and to include collection and storage of clinical specimens as part of that endeavour. Not only is such information essential for immediate disease control, but it will also make possible future research with analytical or laboratory methods yet to be devised. ■

Peter M. Strebel and Stephen L. Cochi are at the National Immunization Program, Centers for Disease Control and Prevention, Atlanta, Georgia 30333, USA.

e-mail: pstrebel@cdc.gov

1. Anderson, R. M. & May, R. M. *Nature* **318**, 323–329 (1985).
2. Grenfell, B. T., Björnstad, O. N. & Kappey, J. *Nature* **414**, 716–723 (2001).
3. de Quadros, C. A. *et al.* *J. Am. Med. Assoc.* **275**, 224–229 (1996).
4. Hersh, B. S., Tambini, G., Nogueira, A. C., Carrasco, P. & de Quadros, C. A. *Lancet* **355**, 1943–1948 (2000).
5. World Health Organization *Weekly Epidemiol. Rec.* **74**, 429–434 (1999).
6. Gay, N. *et al.* *Commun. Dis. Rep. CDR Rev.* **7**, R17–R21 (1997).
7. Orenstein, W. A. *et al.* *Am. J. Pub. Health* **90**, 1521–1525 (2000).
8. Aylward, R. B., Clements, J. & Olive, J. M. *Int. J. Epidemiol.* **26**, 662–669 (1997).
9. WHO Inter-country Office for Eastern Africa. Summary of presentations, reports and recommendations of the East Africa measles strategy clarification meeting, Nairobi, Kenya. World Health Organization, 11–13 October, 1999.
10. <http://www.nt.who.int/vaccines/globalsummary/timeseries/tsincidenceamea.htm>
11. <http://www.unpf.org/swp/2001/english/tables.html>

High-temperature superconductivity

Charged with smuggling heat

Kamran Behnia

Good conductors of heat are usually good at conducting electricity. So the discovery that electrons in a superconductor can carry an unauthorized amount of heat at low temperatures raises many questions.

Copper oxide compounds are noteworthy for the relatively high temperatures (up to 160 K) at which they lose their resistance to electrical currents and become superconducting. This was discovered by Bednorz and Müller¹ in 1986, and gave birth to the field of 'high-temperature' superconductivity. High temperatures aside, the unconventional nature of the superconducting mechanism in these materials has puzzled researchers ever since.

In 1986, few would have guessed that a shock wave produced by that scientific earthquake would reach an old realm of solid-state physics 15 years later and shake one of its most fundamental concepts. But on page 711 of this issue, Hill *et al.*² report that, at low temperatures (close to absolute zero, or -273°C), a copper oxide material in which the superconducting state has been suppressed violates the Wiedemann–Franz law. This universal law relates two basic properties: the ease with which a solid conducts heat and charge. As this is the first time a material has been found that deviates from this law, it reveals, yet again, that high-temperature superconductors are far from being understood.

Almost 150 years ago, Wiedemann and Franz discovered a remarkable correlation between the thermal and electrical conductivities of various metals — good conductors of electricity are also efficient conductors of heat. At room temperature the ratio of the two conductivities was found to be very similar for a broad range of metals. Several decades and a scientific revolution later, this ratio was linked to two fundamental constants: the quantum of electrical charge, e , and the 'quantum' of entropy, k_B (better known as the Boltzmann constant). Roughly speaking, entropy measures the disorder of electrons and is the microscopic origin of 'heat'.

The fundamental mechanism behind the Wiedemann–Franz correlation is easily grasped. The propagation of electrons in a crystal is impeded by the presence of any imperfection (impurities or defects). So there is always a maximum finite distance that an electron can travel before being scattered. Because the electron carries both heat and charge, scattering will affect thermal and electrical conductivities in the same way. This is strictly true only if the scattering is

'elastic' — that is, when the collision is not associated with any change in the particle's energy. This means the Wiedemann–Franz law should always hold true at a temperature of absolute zero when all scattering becomes elastic. As the temperature tends to zero, the ratio of thermal conductivity to electrical conductivity divided by temperature tends to a constant value. This statement is true for all metals tested to date.

The twentieth century saw the development of the 'band structure' theory of electron behaviour in solids, which treats electrons as individual free entities. Surprisingly, this theory — which ignores the strong electrostatic repulsion between the electrons — has been successful in describing the main electronic properties of various metals. In 1956, Lev Landau revealed the reason behind its success. He formulated the concept of a 'Fermi liquid', in which the interacting electrons are replaced by 'quasiparticles' that interact but can still be treated as particles (in this case, fermions) with the same spin and charge as a free electron. The only difference between quasiparticles and electrons is that the quasiparticle has a modified mass, which reflects the size of the electronic interactions.

It is hard to exaggerate the fecundity of Landau's theory. Most metallic elements and compounds are Fermi liquids. There are even exotic metals — in which strong interactions between electrons create Landau quasiparticles several hundred times heavier than a free electron — that fit within the band-structure theory. But do other types of strongly interacting electronic systems exist? In other words, are Fermi liquids only one of many correlated-electron systems found in nature? This question has been haunting condensed-matter physicists for the past 30 years.

The copper oxides challenged the Fermi-liquid picture almost as soon as they were discovered. And we are still far from understanding the elementary excitations (akin to quasiparticles) found in the metallic ('normal') state of these compounds. These superconductors have unusual origins. The parent compound is a Mott insulator, which should be a conductor according to the standard band-structure theory, but fails to conduct because of strong repulsive interactions between the electrons. A metallic state can be created in a Mott insulator by injecting it with conducting electrons — for example, by altering its chemical structure (chemical doping). This is how a copper oxide insulator is turned into a copper oxide superconductor.

But the Fermi-liquid picture cannot easily explain the metallic state of a doped Mott insulator. For example, angle-resolved photoemission data indicate that Landau quasiparticles appear only in the superconducting state³. These and other experimental findings have led to alternative descriptions

of the elementary excitations in these systems. One recurrent feature of these 'non-Fermi-liquid' theories is the separation of two of the most fundamental properties of an electron: its 'spin' and its charge. In the words of P. W. Anderson, "the electron falls apart" in such a way that different excitations are responsible for carrying its spin and charge. But this idea has been difficult to check experimentally because there is no established method for directly probing transportation of the electron spin.

Hill and colleagues² approach this mystery from a new angle. They investigate the transport of heat in the normal metallic state of a copper oxide superconductor at millikelvin temperatures. In all superconductors, whether low- or high-temperature, the zero-resistance state can be destroyed by applying a high enough magnetic field (the size of the field is determined by the superconducting transition temperature). The copper oxides usually require fields that are too high to achieve experimentally, but the compound studied by Hill *et al.* has a transition temperature of just 20 K, so superconductivity can be removed by applying a much smaller magnetic field. Hill *et al.* measure the thermal and electrical conductivity of the 'normal' state of their compound (at low temperatures and under applied fields), and discover a clear departure from the Wiedemann–Franz law. They show that there is no correlation between thermal and electrical conductivity at very low temperature. Instead, the heat conductivity is consistently greater, although the excess suddenly vanishes below 0.3 K. Because the Wiedemann–Franz law is a natural consequence of the Fermi-liquid picture, this spectacular violation has immediate consequences for understanding the elementary excitations in copper oxides. Taken at its face value, it means that charge and heat are not carried by the same type of electronic excitations, so there may indeed be some sort of spin–charge separation in these materials.

This is really just the beginning. We need to repeat the measurement on superconductors with different amounts of chemical doping to determine exactly when the Fermi-liquid picture breaks down. Moreover, studies of other compounds at higher magnetic fields may indicate whether the excess heat flow is associated with electron spin, as suspected. Most pieces of this huge puzzle of modern condensed-matter physics are still waiting to be put into place. ■

Kamran Behnia is in the Laboratoire de Physique Quantique, Ecole Supérieure de Physique et de Chimie Industrielles, 10 rue Vauquelin, 75005 Paris, France.

e-mail: kamran.behnia@espci.fr

1. Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64**, 189–193 (1986).
2. Hill, R. W., Proust, C., Taillefer, L., Fournier, P. & Greene, R. L. *Nature* **414**, 711–715 (2001).
3. Kaminsky, A. *et al. Phys. Rev. Lett.* **84**, 1788–1791 (2000).



100 YEARS AGO

The practicability of effecting the purification of town sewage on the large scale by bacterial agency has now been abundantly proved. The process has passed beyond the experimental stage, and must now be acknowledged as the only method which can convert the putrescible matter of sewage on the large scale into inoffensive and harmless substances. Accordingly all trustworthy information respecting the results which have been arrived at from the lengthy experimental trials, and from the application of these results on the large scale, will be welcome to public sanitary authorities, and perhaps even still more acceptable to the professional advisers of these bodies. The treatise under review has been written by one who has carefully watched the progress, and who has had a long and varied experience, of bacterial treatment. The book is, therefore, undoubtedly worthy of careful perusal and consideration by those who are responsible for disposing of the sewage from houses, villages or towns.

From *Nature* 12 December 1901.

50 YEARS AGO

Teletherapy units using radium are limited in usefulness by the low radiation intensities produced by the small amounts of radium which can be used. To secure an adequate dosage-rate, the distance between the source and the tumour cannot be more than a few centimetres, and therefore the dose delivered to the skin lying between the source and the tumour is much higher than that delivered to the tumour. The dose-rate below the surface, expressed as a percentage of the dose-rate at the skin, decreases very rapidly with increasing depth. Thus the percentage depth-dose is influenced primarily by the inverse square law, and one of the chief advantages of high-energy radiation, namely, its small attenuation by the tissue between the source and the tumour, is not realized. Any attempt to obtain an improvement in the depth-dose by increasing the amount of radium, and correspondingly improving the ratio between the source-to-tumour distance and the source-to-skin distance, is limited by the high cost of radium, and by the required increase in the volume of the source. If the diameter of the source is increased, it is harder to get a well-defined beam; if the thickness is increased, much of the radiation is lost by absorption within the source.

From *Nature* 15 December 1951.

Biomaterials

Sacrificial bonds heal bone

John Currey

The toughness of bone is usually attributed to its collagen, but how does it work? New evidence shows that molecular bonds can temporarily sacrifice themselves to absorb impacts.

Bone's 'toughness' — its ability to absorb impacts and resist the weakening effects of small scratches and holes — is almost as vital to its function as its 'stiffness'. The stiffness of bone, which provides skeletal rigidity, is reasonably well understood: it is produced by tiny mineral crystals that permeate its organic material. Stiffness varies considerably according to the relative amounts of mineral, organic material and water. However, the origin of toughness is still somewhat mysterious. In an article on page 773 of this issue, Thompson and colleagues¹ suggest that collagen — the main organic constituent of bone — has 'sacrificial bonds' that are broken upon loading without significantly harming the bone. Bone's toughness comes from the mechanical work required to overcome these bonds. Some of these sacrificial bonds reform when the force is removed, allowing bone to repair itself to some extent.

A tough material such as bone requires much mechanical work before it fractures. Bone's force–extension curve (Fig. 1) gives a graphical representation of this toughness. Up to the 'yield' point — the range in which bone normally functions — a significant amount of force is required to cause extension. After yield, the curve flattens out, so elongation requires very little extra force. Atomic force microscopy was originally developed to produce highly detailed images of surfaces, but it can also be used to produce force–extension curves for single protein molecules, such as collagen. The protein is stretched by a diamond on the end of a cantilever so that the force required, and the distance that the protein is stretched, can be measured. In addition, the diamond can be pressed into the surface of hard materials such as bone, and the force of the elastic recoil can be measured to determine the mechanical properties at that point in the surface².

Thompson and colleagues¹ first tried this technique on collagen molecules spread on a glass sheet. They found that the force–extension curve, instead of rising smoothly, was saw-toothed with periodic drops. This meant that the collagen kept on extending, requiring significant work in the process, but was never subjected to forces large enough to cause rupture. They also found that less work was needed to extend the collagen on subsequent stretches but that

the work needed gradually recovered with time, although never completely. Thompson and colleagues attributed these saw-toothed patterns to the breaking of sacrificial bonds — bond breakage causing a sudden decrease in the force required — which gradually reformed when the force was removed. The experiment was repeated on molecules exposed on the surface of polished bone, and similar effects were found. In particular, the time course of the recovery in work required was almost the same. These saw-toothed patterns are not visible in the analogous region of the force–extension curve for bone (the flat part of the curve) owing to the inability to obtain such high resolution on whole specimens.

A clue to the nature of the sacrificial bonds came when the authors soaked the collagen or bone in different solutions. More work was required on subsequent stretches when a solution containing divalent calcium ions was used than when the solution contained monovalent sodium ions. Indeed, the specimens showed little saw-tooth behaviour in the sodium-containing solution. As divalent, but not monovalent, ions can link molecules together, this suggests that some kind of ionic bond between charged residues in collagen might be responsible for reconnecting the individual collagen molecules to each other. Furthermore, previous work from this group has shown similar results for the organic material between the little plates of calcium carbonate in mother of pearl³, so it could be that this method of toughening is widespread among calcified tissues.

These findings are potentially important for understanding the postyield behaviour of bone (Fig. 1). Nevertheless, there is still a considerable element of speculation in all this. Thompson *et al.* cannot be sure that they are stretching collagen molecules in their experiments with bone, although it is likely that they are, given the similar saw-toothed curves. In addition, they have no means of knowing whether they are stretching a single molecule or more than one, which might confound the interpretation of the saw-toothed curve. Finally, it is curious that they find very similar amounts of work to be required when pulling on collagen and bone. In bone, collagen is bound rather tightly to the mineral, so the ability to stretch free collagen molecules seems unlikely.

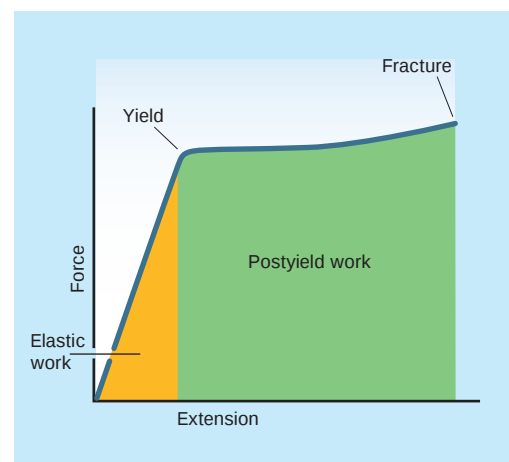


Figure 1 Force–extension curve for bone. Bone is elastic up to the yield point and so it can return to its original state when unloaded. This is the range in which bone normally functions. At yield, and after, damage occurs such that bone remains distorted upon unloading and is less stiff than before; that is, it requires a smaller force to produce the same extension. Thompson and colleagues¹ suggest that this reduced stiffness is due to the temporary breakage of sacrificial bonds within the bone's collagen molecules, which begins after the bone has passed its yield point.

Some researchers think that the postyield behaviour of bone is caused by extensive microcracking^{4,5}. When bone undergoes postyield extensions, innumerable microcracks begin, spread slightly, but then come to a halt. So this flat part of the curve is thought to be caused by the reduced stiffness induced by the microcracks. In other words, bone is damaged by these microcracks, but remains coherent until fracture. Microcrack-induced elasticity can adequately explain the behaviour of bone in qualitative terms. However, at present it is impossible to assign numbers to this effect and the qualitative explanation might not, in the end, give the right quantitative answers.

Others, including of course Thompson *et al.*, think that the postyield behaviour of bone is best considered at a more molecular level than microcracking — at the level of bonding of mineral crystals to the organic matrix and of molecular interactions within collagen itself. Whether this latter mechanism is involved in preventing the spread of microcracks is unknown; if it is, it would be fascinating to see the two theories reconciled. ■

John Currey is in the Department of Biology, University of York, York YO10 5YW, UK.
e-mail: jdc1@york.ac.uk

1. Thompson, J. B. *et al.* *Nature* **414**, 773–776 (2001).
2. Oliver, W. C. & Pharr, G. M. *J. Materials Res.* **7**, 1564–1583 (1992).
3. Smith, B. L. *et al.* *Nature* **399**, 761–763 (1996).
4. Zioupos, P. & Currey, J. D. *J. Materials Sci.* **29**, 978–986 (1994).
5. Vashishth, D., Tanner, K. E. & Bonfield, W. *J. Biomech.* **33**, 1169–1174 (2000).

Space physics

Rhythms of the auroral dance

Patrick T. Newell

Earlier this year, the four satellites of the Cluster mission passed through part of the electric circuit that causes aurorae. Their observations support the view that intense aurorae form in regions largely devoid of electrons.

Early in the space age, measurements from rocket flights established that bright auroral displays (Fig. 1) are associated with beams of accelerated electrons from space¹. Dynamic displays that are visible with the unaided eye are known as discrete aurorae. Ground-based research on auroral forms and motions, using all-sky cameras, suggested that the electrons are accelerated by electric fields aligned along the direction of the Earth's magnetic field lines². Later, more sophisticated satellite measurements established the existence of

such electric fields³. But it has always been difficult to reconcile these quasi-steady electric fields with standard plasma theory, which predicts that cold background electrons in the ionosphere should quickly move in the direction opposite to the electric field and cancel it out.

So we have a long-running mystery. Understanding when and how these magnetic-field-aligned electric circuits are created is key to explaining how aurorae form, and one important question is the timescale over which the fields evolve. On page 724 of

this issue⁴, Marklund and colleagues report on the first multi-satellite encounter with the auroral circuit, and the resulting observations that bear on this question.

There are severe limitations on the measurements possible from a single satellite. In particular, when the satellite records a change, it is usually unclear whether the structure of the auroral electric circuit is really changing, or whether the satellite is simply moving to a different part of it. This spatial-temporal ambiguity can be addressed by using several coordinated satellites, preferably with identical instruments. The first such mission to be flown is the European Space Agency's Cluster, which consists of four tightly grouped satellites, whimsically named after Latin dances: Rumba, Salsa, Samba and Tango. The Cluster satellites had a perigee (point of orbit closest to Earth) well above the expected electron acceleration region, the site of the important physics. So the satellites' encounter with the region was not predicted, and is all the more exciting for that.

Figure 1 of Marklund and colleagues' paper (page 724) shows the overall auroral circuit, in which current flows towards Earth from space, through the magnetosphere and the ionosphere, and then out into space again. Figure 2 here depicts the two acceleration regions (downward and return) in more detail. These regions include currents, driven by magnetospheric processes, and the mysterious electric fields that accelerate electrons to the high energies needed to create aurorae. Surveys from single-satellite missions have shown⁵ that auroral electric fields rarely occur above an altitude of 15,000 km. Yet on 14 January 2001 Cluster encountered just such a rarity: Rumba, Salsa, Samba and Tango traversed a portion of the auroral acceleration structure at an altitude of 22,000 km, passing through it like beads on a string.

This chance encounter was actually with the return current region, shown in Fig. 2b. The satellites' most notable finding was that the electric fields evolved over a characteristic timescale of 200 seconds. Given the observed strength of the auroral return current, this would be about the time needed to completely evacuate the geomagnetic field line of electrons. Electric currents must form a closed loop (in the absence of a build-up of charge in space), meaning that the return current should exhibit the same behaviour as the downward current above the aurora itself. These findings support the interpretation that intense aurorae are associated with a reduction of the background electrons in the ionosphere to near-zero densities.

The results from Cluster's fortuitous encounter with auroral acceleration structures agree well with other findings, such as the tendency for discrete aurorae to form in darkness, rather than sunlight⁶ (the reason is that ultraviolet light from the Sun creates free

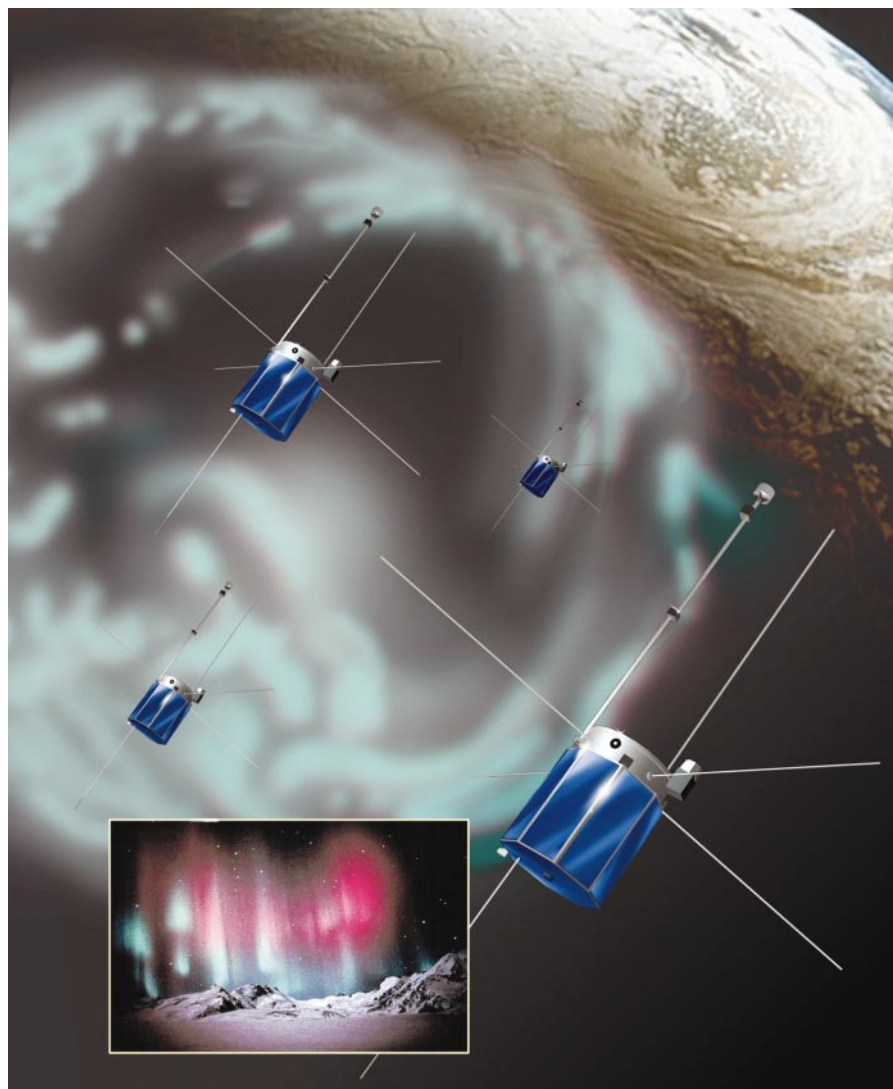


Figure 1 Auroral displays seen from space and (inset) from Earth. The main image is an artist's impression of what we could see in the future: a mission along the lines of Cluster, but flown at altitudes more suitable for auroral observation. Several satellites with identical instruments should traverse the aurora simultaneously. (Main image courtesy of B. Mauk, JHU/APL.)

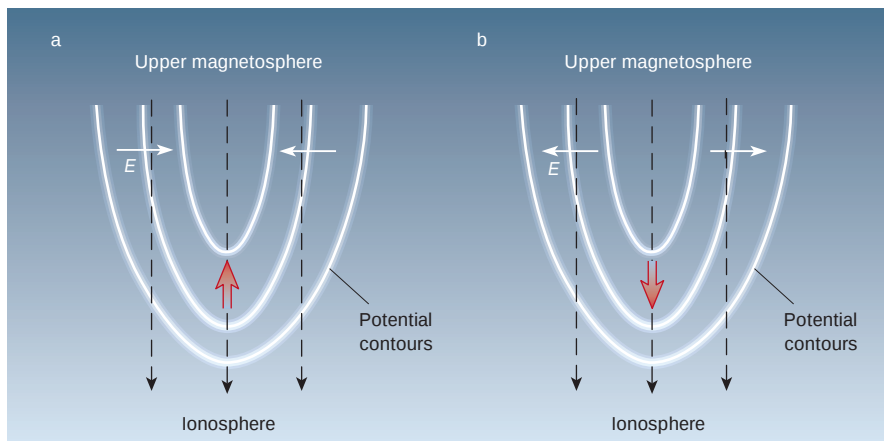


Figure 2 The auroral circuit. (See also Fig. 1 on page 724.) a, Converging electric fields occur in regions of currents leaving the ionosphere and create intense aurorae. Here the electrons (not shown) are being accelerated downwards, so current flow is upwards (because electrons are negatively charged, current flows in the opposite direction to electrons). b, These currents are 'closed' by the region of auroral return current, which has been observed by satellites of the Cluster mission and is described by Marklund *et al.*⁴. Here electrons are being accelerated upwards, so current flows downwards. Black dashed lines show Earth's magnetic field; red arrows indicate the auroral currents; *E*, the accelerating electric fields. Acceleration occurs because charged particles such as electrons gain or lose energy when crossing potential contours (lines of constant electric potential).

electrons in the ionosphere, by separating them from neutral atmospheric atoms). The definitiveness of Marklund and colleagues' conclusions is limited by the fact that the observation was necessarily of an auroral oddity (the auroral structure was unusually high in altitude, or it would not have been detected), and because only the auroral return current and associated electric field were observed. Whether flights through a typical aurora, preferably through the downward acceleration region, would show the same behaviour cannot be known for sure. A dedicated multi-satellite auroral mission, such as shown in Fig. 1, is needed for a final verdict.

It is clear, though, that Marklund and colleagues' paper represents a first look at the future of auroral research. The spatial and temporal ambiguities of single-satellite

measurements have hampered progress for years, and nearly all auroral scientists are eager to see a Cluster-like configuration flown at altitudes typical of the aurora. In the meantime, the rich data set provided by Marklund *et al.* will be scrutinized thoroughly to see whether inferences made from single-satellite measurements hold up. ■

Patrick T. Newell is at The Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, Maryland 20723, USA.
e-mail: patrick.newell@jhuapl.edu

1. McIlwain, C. E. *J. Geophys. Res.* **65**, 2727–2747 (1960).
2. Hallinan, T. J. & Davis, T. N. *Planet. Space Sci.* **18**, 1735–1744 (1970).
3. Mozer, F. S. *et al. Space Sci. Rev.* **27**, 155–213 (1980).
4. Marklund, G. T. *et al. Nature* **414**, 724–727 (2001).
5. Reiff, P. H. *et al. in Auroral Plasma Dynamics* (ed. Lysak, R. L.) 143–154 (Am. Geophys. Union, Washington, DC, 1993).
6. Newell, P. T. *et al. Nature* **381**, 766–767 (1996).

Immunology

Bug detectors

Tsuneyasu Kaisho and Shizuo Akira

Insects and mammals are thought to have similar 'innate' immune responses to infectious microorganisms. But there are differences, and even more now emerge from studies of how insects detect bacteria.

To respond rapidly to invading microorganisms, animals call on their innate immune systems. These consist of microbe-detecting molecules, which are connected to cascades of signalling proteins that in turn activate immune cells. Innate immunity has been particularly well studied in insects and mammals, and these animals have many similar molecules and mechanisms. For example, both rely on certain

cell-surface proteins^{1,2} — from the Toll family in fruitflies, and the related Toll-like receptor (TLR) family in mammals. But there are crucial differences, too: in mammals, for instance, the TLR proteins detect microbes directly by recognizing the molecular patterns that the microbes display, whereas fruitfly Toll proteins do not seem to do this. So what are the bug-detecting molecules in insects? On page 756 of this

issue, Michel and colleagues³ provide an answer — at least for some types of microbe.

The Toll and TLR proteins are strikingly alike in structure, and have many other similarities. For example, it is assumed that different members of the two protein families are used to respond to different types of pathogen or pathogen component. In mammals, TLR2 and TLR4 are essential for the immune response to, respectively, peptidoglycan (the main constituent of the cell walls of 'Gram-positive' bacteria) and lipopolysaccharides (the major cell-wall component in 'Gram-negative' bacteria). In fruitflies (*Drosophila melanogaster*), a signalling pathway involving Toll is required for the response to Gram-positive bacteria and fungi, but Toll is not needed for the response to Gram-negative bacteria (although it is not yet known which Toll-family member is required). Moreover, many of the signalling molecules that are activated in response to Tolls and TLRs are similar.

It is becoming clear, however, that there are also many differences between the Toll and TLR systems (Fig. 1, overleaf). In mammals, Gram-positive and Gram-negative bacteria activate a common intracellular signalling cascade, albeit through different TLR proteins. But in insects, Gram-positive bacteria and fungi activate one pathway, whereas Gram-negative bacteria activate another. Also, mammals use many different TLRs that appear to detect different microbial components directly. But so far only one of the nine Toll-family members has been definitively shown to be involved in *Drosophila* innate immunity, and this protein (Toll) does not detect molecular patterns on microbes. In fact, it is not activated directly by pathogen-derived molecules at all, but by another *Drosophila* protein, Spätzle. After fruitflies are infected, cascades of protein-cleaving enzymes called serine proteases are somehow switched on, and these process Spätzle into the active form that can work on Toll.

So which proteins actually detect pathogens in *Drosophila*? To find out, Michel *et al.*³ carried out screens for mutant fruitflies that cannot produce the antimicrobial peptide Drosomycin (which is normally generated once the Toll pathway has been activated by Gram-positive bacteria or fungi). The authors identified such a mutant, which they named *semmelweis* (*seml*). *Drosophila* with the *seml* mutation can no longer activate Drosomycin in response to Gram-positive bacteria, but can do so in response to fungi. So it seems that, although both types of pathogen work through Toll to connect to the same intracellular signalling pathway, they are initially detected in different ways.

Next, Michel *et al.* tracked down the precise gene that is affected by the *seml* mutation, and found that it encodes a known peptidoglycan-recognition protein, PGRP-SA. The *seml* mutation results in a cysteine

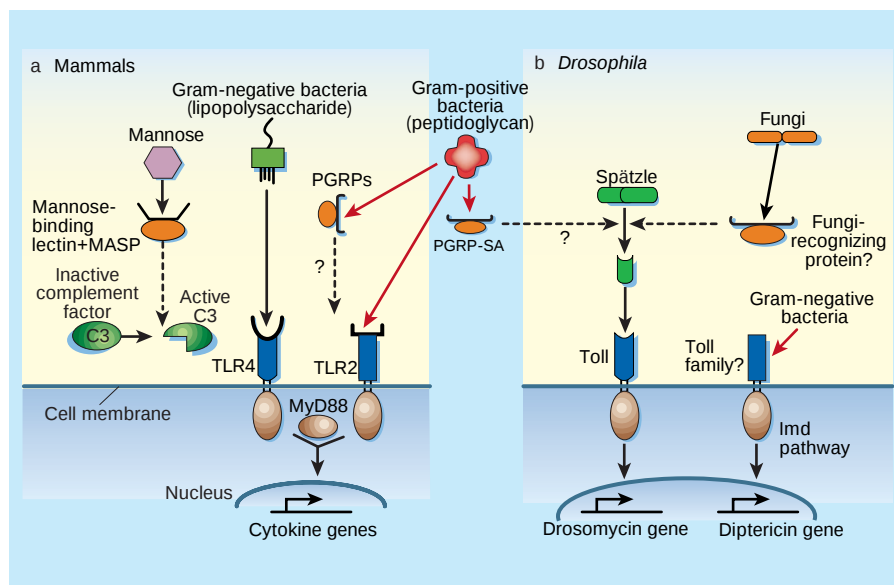


Figure 1 Innate immunity in insects and mammals. The fruitfly Toll and mammalian TLR proteins have similar structures but many different properties. a, In mammals, microbial components (such as peptidoglycans in Gram-positive and lipopolysaccharides in Gram-negative bacteria) are detected directly by different TLRs, which activate a common signalling pathway (involving the protein MyD88). b, In *Drosophila*, the response to fungi and Gram-positive bacteria is mediated through Toll, and the response to Gram-negative bacteria through another Toll-family protein. But different intracellular signalling pathways are induced. Also, Toll is not activated by pathogens directly, but through another *Drosophila* protein, Spätzle. Michel *et al.*³ now show that peptidoglycans are detected by PGRP-SA in the blood. This protein presumably activates cascades of protease enzymes (dashed arrows) that in turn lead to the production of active Spätzle. PGRP-SA is not needed for antifungal immunity in insects. Mammals have different sensors in the blood (including the mannose-binding lectin) that activate protein cascades. It remains unknown whether mammalian PGRPs can lead to the activation of TLRs or protease cascades.

amino acid being replaced by a tyrosine in the protein; this would presumably alter the protein's three-dimensional structure and hence its function. It seems likely that, in normal circumstances, when the wild-type PGRP-SA recognizes peptidoglycan in the cell wall of Gram-positive bacteria, as yet unknown serine proteases are triggered, leading to the production of active Spätzle, which can in turn switch on Toll.

From an evolutionary point of view, this is all very interesting. Michel *et al.*³ show that the *Drosophila* PGRP-SA is found in the haemolymph, an insect fluid that is similar to mammalian blood and lymph. In horseshoe crabs, too, innate immunity is triggered through the recognition of specific microbial patterns by molecular sensors in the equivalent of the blood circulation⁴. Here, lipopolysaccharides and (1→3)- β -glucan — components of the cell wall in Gram-negative bacteria and fungi, respectively — activate two precursors of serine proteases, causing them to cleave themselves and then trigger signalling cascades that lead to blood coagulation. The resulting gel, coagulin, engulfs or immobilizes the microorganisms.

Mammals, too, have a related system, involving a protein known as mannose-binding lectin⁵. This is found in blood plasma, and binds to microbial carbohydrate structures that contain mannose or *N*-acetyl-

glucosamine groups. The protein then activates MASPs, a group of serine proteases, and thereby molecules of the 'complement' system, resulting in increased engulfment or lysis of microorganisms. Another group of lectins, the ficolins, also recognize carbohydrates containing *N*-acetylglucosamine, and activate complement factors via the MASPs. And a group of invertebrates

(ascidians or sea squirts) also use lectins to activate complement factors. So the ability to detect microbes in the circulation seems to have been conserved during evolution.

There are even hints that — despite its usual cell-surface distribution — mammalian TLR4 can recognize host molecules such as the protein fibrinogen, rather than pathogen-derived ones, in the blood system. Moreover, lipopolysaccharide- and peptidoglycan-binding proteins have been detected in mammalian blood serum. Nevertheless, it seems that the mammalian TLR and *Drosophila* Toll systems have evolved in different directions. There is no evidence to link the microbe-recognition proteins in mammalian serum with protease cascades that could activate TLRs. And there is the crucial distinction that — unlike Toll — TLR4 and TLR9 recognize their ligands directly.

Another issue raised by the new results³ is that both Gram-positive bacteria and fungi lead to the activation of Toll, yet only Gram-positive bacteria are detected by PGRP-SA. So it remains unknown how insects detect fungi, although it can be assumed that any fungi-recognizing proteins would also lead to cleavage of Spätzle. Further genetic screens should reveal the complete signalling pathways, as well as the true extent of the similarity between mammalian and insect innate immunity.

Tsuneyasu Kaisho and Shizuo Akira are in the Department of Host Defense, Research Institute for Microbial Diseases, Osaka University, Osaka 565-0871, and SORST, Japan Science and Technology Corporation, Japan.
e-mail: sakira@biken.osaka-u.ac.jp

1. Imler, J. L. & Hoffmann, J. A. *Trends Cell Biol.* **11**, 304–311 (2001).
2. Akira, S., Takeda, K. & Kaisho, T. *Nature Immunol.* **2**, 675–680 (2001).
3. Michel, T., Reichhart, J.-M., Hoffmann, J. A. & Royet, J. *Nature* **414**, 756–759 (2001).
4. Muta, T. & Iwanaga, S. *Curr. Opin. Immunol.* **8**, 41–47 (1996).
5. Gadjeva, M. *et al.* *Curr. Opin. Immunol.* **13**, 74–78 (2001).

Chemistry

An end to the protection racket

David Gani

To modify a specific site on a molecule that has several similar sites, organic chemists have had to use cumbersome 'protecting groups'. New peptide catalysts with remarkable selectivities offer a protection-free alternative.

Frustration is a familiar feeling for organic chemists. The synthesis of complex molecules presents innumerable challenges and leaves the chemist envious of the ease with which nature performs such tasks. Cells contain countless enzymes — biological catalysts that oversee the efficient conversion of substrates to products. But the synthetic chemist rarely finds an enzyme that will catalyse a specific chemical reaction in the laboratory.

Moreover, the downside of high enzyme selectivity is limited tolerance for substrates that 'fit' poorly. So coaxing a highly evolved enzyme to process even closely related substrates is usually futile. Tinkering with the structure of enzymes to broaden tolerance is possible, but such 'protein engineering' is difficult, and requires detailed information about how the protein's active site recognizes substrates at the molecular level. But, as Sculimbrene and Miller¹ reveal in the

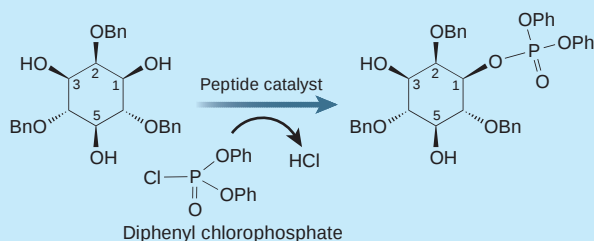


Figure 1 Synthetic peptide catalyst in action. Sculimbrene and Miller¹ have developed a peptide catalyst that can selectively add a phosphate group to inositol-derived 1,3,5-triol, which contains three similar hydroxyl (OH) sites. Diphenyl chlorophosphate is used as a phosphoryl donor. Because the triol substrate is symmetrical (the plane of symmetry runs through carbon atoms 2 and 5), it does not exist as two distinct mirror images (enantiomers). The product, by contrast, is a single enantiomer: only one of the three hydroxyl groups is phosphorylated. The other enantiomer (not shown) would be phosphorylated at carbon atom 3. This remarkable selectivity approaches that achieved by biological catalysts. In a subsequent step, the authors removed the phenyl (Ph) groups.

Journal of the American Chemical Society, there is a way forward. The authors show that selective, enzyme-like catalysts can be obtained simply, starting with the synthesis of libraries of potential catalysts. They have identified a catalyst that selectively modifies a single site on a substrate that has three similar reaction sites.

The synthetic peptide created by Sculimbrene and Miller¹ is a kinase mimic. In the cell, kinases typically catalyse the transfer of a phosphoryl group (PO_3^{2-}) from adenosine triphosphate (ATP) to a hydroxyl group (OH) of an alcohol to produce a phosphate ester. Like most enzymes, kinases are highly selective. They phosphorylate only one hydroxyl group in a site-selective manner, even when there are many similar hydroxyl sites in the substrate molecule.

This selectivity arises from exceptional substrate recognition by the enzyme, and is crucial for maintaining the metabolic balance in the cell. Indeed, kinases and their dephosphorylating counterparts, phosphatases, modulate the activity of numerous cellular pathways by making sure the right number of key molecules are phosphorylated. The negatively charged phosphate esters act as switches to turn a

specific activity, such as receptor binding, on or off.

To attain similar selectivity, chemists have, until now, relied mainly on two strategies. One uses limited amounts of reagent but inevitably gives only low yields of the desired product. This method is direct (requiring just one chemical step), but defers the real problem to the separation of the desired material from unwanted side products, which is wasteful and very inefficient. The other strategy is to 'protect' all of the similar sites, except those requiring transformation, by introducing a blocking group.

In this process, the protected molecule can then be exposed to excess reagent to modify the unprotected site (or sites) as desired. Although the yield of the required products is good, this method has many drawbacks. Selectively blocking only some sites is not always possible. In addition, this method is undesirable for large-scale processing, because of the extra chemical steps required to both introduce and then remove the bulky blocking groups—which are often larger than the substrate itself. This strategy generates an enormous amount of waste, and is not environmentally friendly.

The peptide developed by Sculimbrene

and Miller¹ scores highly in terms of selectivity. Their catalyst delivers a phosphoryl group—in a biomimetic fashion²—to just one of the three similar hydroxyl sites on an inositol derivative. The product is a specific inositol 1-phosphate compound (Fig. 1). Indeed, the product is a single enantiomer (one of two possible 'chiral', mirror-image versions of the same structure), even though the substrate is not itself an enantiomer. So how does their catalyst work and why is it so selective?

In earlier work^{3,4}, the authors found that several octapeptides (peptides eight amino acids long) could selectively acylate a 1:1 mixture of alcohol enantiomers. In each case, the peptide catalysed the selective transfer of an acetyl group (CH_3CO) to an oxygen atom on one of the two alcohol enantiomers. They were able to fine-tune the reaction to achieve a 50:1 ratio for one product enantiomer over the other (Fig. 2). Although exciting, these results were anticipated because the octapeptides themselves were chiral, and so should interact differently with the different substrates. In other words, a specific catalyst enantiomer preferentially gave a specific product enantiomer.

In their latest study¹, Sculimbrene and Miller prepared a small library of 39 pentapeptide catalysts to mimic the activity of enzymes in cellular phosphoryl-transfer reactions. The switch from acylation to intrinsically slower phosphorylation chemistry is conceptually insignificant but allows for more selectivity and useful end products. Library members were screened for the ability to selectively phosphorylate one site on an inositol-derived 1,3,5-triol substrate, which has three very similar hydroxyl sites. One of the pentapeptides was highly selective, so the authors resynthesized and purified it for further evaluation. Remarkably, using the pure catalyst resulted in only one of the possible enantiomers of the product. A small amount of 1,3-bisphosphate ester was the only unwanted phosphorylation product.

Together, the various studies by Sculimbrene and Miller demonstrate principles

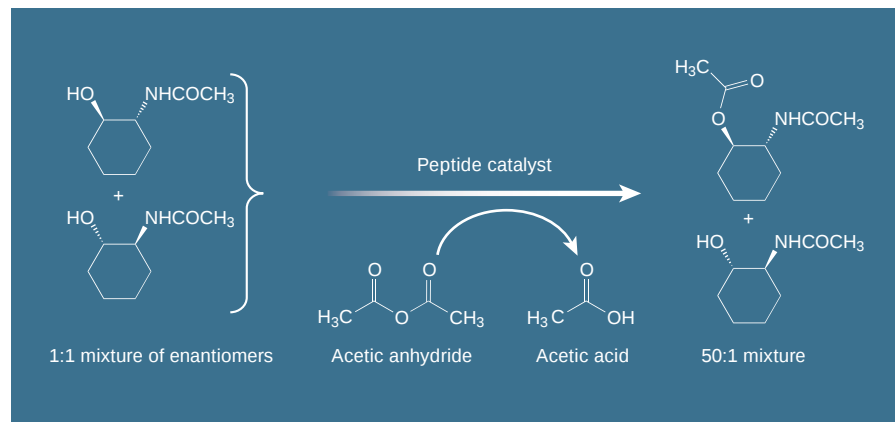


Figure 2 Favouring one reaction over another. Acetylating a 1:1 mixture of alcohol enantiomers using acetic anhydride and a peptide catalyst favours the production of one acylated enantiomer over another⁴. Because the peptide catalysts chosen for this reaction were themselves chiral, the prospects of achieving an enantioselective reaction were enhanced.

that can easily be transferred to many other reaction types and substrates. The findings also provide a boost to clean, green chemistry as the peptide catalysts are cheap, easy to prepare and are not particularly hazardous. Another area of potential, which is not as far along as Sculimbrene and Miller's work, is a group of related poly-leucine-type catalysts, which seem to have promising enantioselectivities⁵. New catalysts of this type could change organic synthesis as much as the catalytic chemistries of Knowles, Nyori and Sharpless, for which they were awarded this year's Nobel prize in chemistry. But, for the peptide-based systems, much more research is needed to understand how the catalyst–substrate complexes form, what their structures are, and how chemists might manipulate their site selectivities. The

efficient and selective synthesis of speciality chemicals, such as drugs and agrochemicals, and other 'biological effect' molecules remains a major industrial challenge. The importance of this area is so profound that it won't be long before we start to have answers to these fundamental questions, and use this information in designing a wider range of exquisite catalysts.

David Gani is in the School of Chemical Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.

e-mail: dgani@chemistry.bham.ac.uk

1. Sculimbrene, B. R. & Miller, S. J. *J. Am. Chem. Soc.* **123**, 10125–10126 (2001).
2. Pirrung, M. C. *Chem. Biol.* **6**, R167–R175 (1999).
3. Jarvo, E. R. *et al.* *J. Am. Chem. Soc.* **121**, 11638–11643 (1999).
4. Copeland, G. T. & Miller, S. J. *J. Am. Chem. Soc.* **123**, 6496–6502 (2001).
5. Bentley, P. A. *et al.* *Chem. Commun.* 1616–1617 (2001).

Medicine

Short cut to disease genes

Alan F. Wright and Veronica Van Heyningen

There are thousands of inherited human diseases. For the most part, the genes that are mutated in these diseases are unknown. But a new approach could save time and effort in identifying the genes responsible.

Finding genes that are implicated in human diseases has never been easy. But when it comes to genetically complex diseases, such as diabetes or psychiatric disorders, the task is particularly hard. The reasons are clear: there can be many, diverse genes involved; their individual effects are weak; and the background noise from non-genetic

influences can be deafening. Moreover, a major bottleneck—even for disorders caused by mutation of a single gene—is 'mapping' the defect to a small enough genomic region for the gene to be identified. Researchers typically sift through tens or hundreds of candidate genes in a mapped region, hoping to find some that are functionally relevant or

strongly expressed in the affected tissue. Writing in *Cell*, Cepko and colleagues¹ suggest that the whole process could be short-circuited by taking a different tack—first finding genes that are preferentially expressed in a given tissue, and then looking to see if those genes are located in mapped genomic regions that contain unidentified disease genes.

The idea is simple. If an inherited disease preferentially affects a particular tissue, the chance of successfully identifying the responsible gene will be increased by restricting the search to genes that are more highly expressed there than in other tissues (as judged, for example, by the number of messenger RNA copies of those genes). Knowing the tissue of interest is one thing with visual or hearing disorders, but quite another in type 2 diabetes, where a gene influencing disease might be expressed in the pancreas, fat, liver or several tissues. Yet restricting the search to genes expressed in any of these tissues could still save time and effort. Surprisingly, this seemingly obvious approach has never been systematically investigated. Nor has the related possibility that disease genes could be distinguished not only by their relative mRNA abundance in the affected tissue compared with other tissues, but also by their absolute mRNA abundance in the tissue in question.

Cepko and colleagues¹ tested the approach in the retina, focusing on a single cell type known as the rod photoreceptor, which is strongly implicated in retinal diseases. In rodents, photoreceptor precursor cells appear late in development², and groups of new genes are switched on as the precursors become committed to becoming rod cells. Expression of half of the 71 genes known

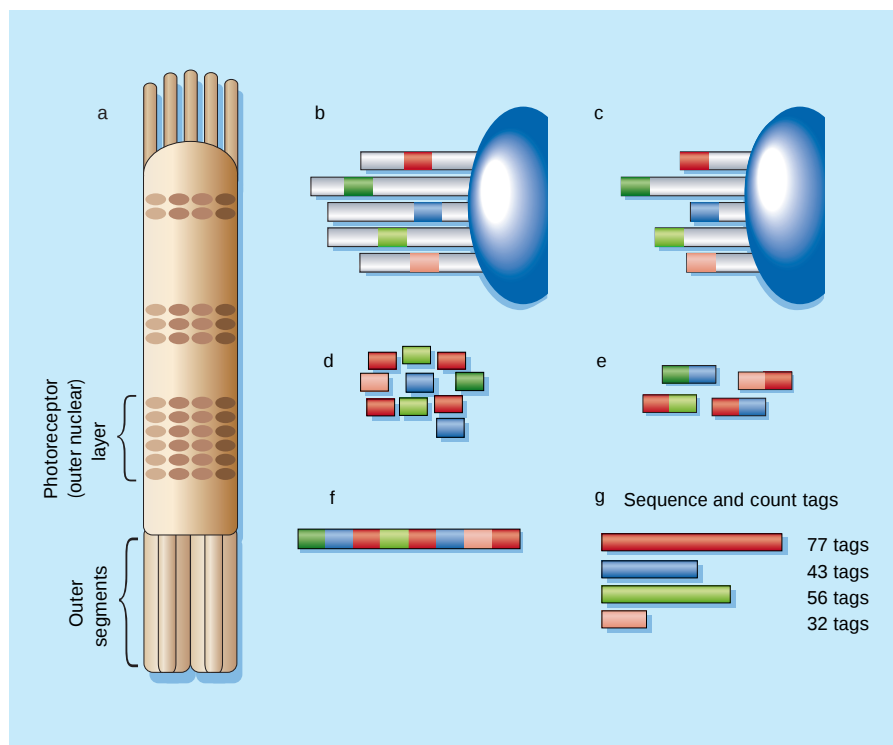


Figure 1 Serial analysis of gene expression (SAGE) in the retina, as carried out by Cepko and colleagues¹. a, Cross-section of the retina, showing the three main layers of neuronal nuclei (ovals). The photoreceptor outer segments are involved in transducing light. In the outer nuclear layer, 97% of cells are photoreceptors. b, Messenger RNAs are isolated from the tissue and captured on beads, where double-stranded DNAs, complementary to the mRNAs, are synthesized. c, The complementary DNAs (cDNAs) are cleaved near their 3' end. d, SAGE tags (coloured blocks, representing short, unique sequence fingerprints) are released by cleavage with another enzyme. e, After splitting the sample into two, tags from each half are ligated together to form unique 'ditags', so that they can be multiplied without introducing bias. f, Ditags are joined together in a linear array and cloned into a sequencing vector so that many different tags can be sequenced at once. g, Sequence analysis allows mRNA abundance to be estimated efficiently^{4,5}.

to cause human retinal diseases is increased (enriched) in rod photoreceptors compared with other tissues^{1,3}. The reason that such a high number of genes (71) causing retinal diseases have been identified is because photoreceptor function has an unusually direct and measurable relationship to the clinical 'phenotype' — visual disability. So it is relatively easy to detect even subtle clinical problems and the genes responsible for them.

This close link between genotype and phenotype made photoreceptors a good model system for Cepko and colleagues¹. There are several techniques for profiling the gene-expression pattern of a cell or tissue. The authors used 'serial analysis of gene expression' (SAGE)^{4,5} to quantify the mRNAs isolated from developing and mature mouse retinas (Fig. 1). Each SAGE 'library' contains large numbers of 'tags' — short sequences from the so-called 3' end of an mRNA, most of which are unique to that mRNA. A relatively unbiased estimate of the abundance of the mRNA type is obtained by simply counting tags. The authors counted over 50,000 tags per library, and identified 70–80% of the different mRNAs expressed at each retinal stage⁶. Using four criteria, they then picked out those mRNAs that are both moderately abundant and enriched in rods.

The authors examined 1,119 different tags that are abundant in the retina, and found that about half satisfied at least two of the criteria for rod enrichment¹ (although two-thirds were also expressed outside the retina). Next, they looked back at 22 retinal-disease genes that had been mapped and identified, and showed that 11 fell into the 'increased expression in the rod' class. Extending this to all identified retinal-disease genes gave a similar result, with 33 of 71 showing rod-enriched expression. However, many of these disease genes were identified by the more traditional map-search approach precisely because they are highly expressed in the retina or affect vision-specific processes.

Cepko and colleagues therefore looked in detail at 237 uncharacterized, moderately abundant genes that show increased expression in rod cells, and identified their human counterparts and map positions. They found that 86 of these genes map to 37 regions containing unidentified retinal-disease genes, suggesting that they are excellent candidates. Subsequent work has already confirmed that mutations in at least one of these genes underlie a retinal disorder (S. Daiger, personal communication). If the same approach holds for other diseases, the advantage is considerable — a 30–100-fold reduction in the number of genes to be screened.

One interesting feature of the authors' results is that these 237 more abundant genes represent a minority (perhaps 10%) of the total number enriched in rods^{6,7}. The majority of mRNAs showing tissue-enriched expression are generally found at much

lower abundance. Yet a large proportion of the abundant class is implicated in retinal diseases. Why should a high absolute abundance correlate with involvement in disease? The answer may lie in the genetic mechanism at work. It could be that strongly expressed genes are associated with 'dominant' diseases, where a 50% loss of normal product or 50% gain of aberrant product has a substantial effect on the phenotype of the tissue in question. In contrast, recessive phenotypes are by definition insensitive to a 50% change in a gene product that is typically less abundant and less likely to gain an abnormal function. In line with this hypothesis, one photoreceptor SAGE library¹ showed median values of roughly 300 and 60 mRNA copies per cell for genes associated with dominant and recessive diseases, respectively.

It also seems probable that a higher proportion of dominant than recessive disease genes have been identified so far, because they are easier to map. But there are clearly many retinal-disease genes still to be mapped; the total number is almost certainly two or three times the present figure of 124 mapped genes (versus the 71 that have been identified; ref. 3). So the apparent excess of abundant mRNAs among disease genes could wane as more recessive mutations are identified. Eventually the low-abundance class of genes may turn out to be numerically more important in disease.

Will the new approach¹ be useful in identifying other disease genes? In genetically complex disorders, the relationship between genetic abnormality and disease is far less direct⁸, and the processes that lead to disease are more diverse and less clear, than in the retina. Yet, in many cases, particular cells and tissues are still strongly implicated. So, because the genomic regions to be searched are often much larger than in simpler disorders, the new approach offers a substantial short cut. The new findings are also of considerable biological interest: the discovery of hundreds of photoreceptor-enriched genes will keep biologists busy for years. More generally, the paper clearly shows the need for more in-depth gene-expression profiles, to help researchers in their quest for the genetic basis of inherited diseases. ■

Alan F. Wright and Veronica Van Heyningen are at the MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. e-mail: alan.wright@hgu.mrc.ac.uk

1. Blackshaw, S., Fraioli, R. E., Furukawa, T. & Cepko, C. L. *Cell* **107**, 579–589 (2001).
2. Cepko, C. L., Austin, C. P., Yang, X., Alexiades, M. & Ezzeddine, D. *Proc. Natl Acad. Sci. USA* **93**, 589–595 (1996).
3. RetNet Retinal Information Network; www.sph.uth.tmc.edu/retnet/
4. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. *Science* **270**, 484–487 (1995).
5. Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. *Trends Genet.* **16**, 423–425 (2000).
6. Velculescu, V. E. *et al. Nature Genet.* **23**, 387–388 (1999).
7. Moreno, J. C. *et al. Genomics* **75**, 70–76 (2001).
8. Wright, A. F. & Hastie, N. *Genome Biol.* **2**, 2007.1–2007.8 (2001).

Daedalus

Reduced bandwidth

World culture is dominated by film and television. The alleged need for streaming video and video-on-demand drives many technical efforts, typically to boost computer transmission. But the resulting increase in bandwidth chokes the 'last mile' of old telephone wire to each viewer. And all this technology still relies on the century-old discovery that repeated still pictures, presented at 24 frames a second or more, give most people the illusion of motion. Meanwhile, sensory psychology has moved on.

Daedalus points out that the human eye slides around a scene, and makes several jumps per second (saccades) in the process. With each jump, large changes go unnoticed. This 'change blindness' allows a watching computer to change the hair colour of a human figure during a saccade, or even to remove it entirely, without detection. A convincing moving image could be presented at only a few frames a second if the frames were made to coincide with the shifts of attention of the human visual system itself.

Cinema and digital cameras, of course, register 24 or more complete frames a second; but they are not the problem. DREADCO physiologists are discovering just how few frames, if chosen to change during a saccade or an eye movement, will be sufficient to give the illusion of continuous movement on a video screen. They are studying viewers' eyes to spot the moments of change blindness. With luck, everybody will react alike. If they do, a few well-chosen frames, certainly fewer than 24, should give the illusion of perfect movement.

But even if we are all different, the technique will still work. Each video screen will have a small camera that detects optical cues from the viewer. Such 'interactive systems' are already being tested. Instead of 24 or more complete frames a second, far fewer will do a much better job. Personal data will be relayed back to the junction where broadband video meets the 'last mile' of telephone wire. Final bandwidth, the bugbear of all video systems, will be minimized.

The new DREADCO technology will only work for computer-transmitted video; the old cinema will still need 24 frames a second. But even if audience members all make different saccade choices, we may have some in common. A computer-controlled projector could select its frames, or show only parts of them, saving wear on the film and enhancing the pleasure of the audience.

David Jones

Laterality in tool manufacture by crows

Neural processing and not ecological factors may influence 'handedness' in these birds.

New Caledonian crows (Fig. 1) fashion tapered tools from either the left or the right edge of the long narrow leaves of pandanus trees or screw pines^{1,2}, which they use to extract invertebrates in rainforest vegetation². Although right-handedness is thought to be uniquely human³, we show here that crows from different localities display a widespread laterality in making their tools, indicating that this behaviour is unlikely to be attributable to local social traditions or ecological factors. To our knowledge, this is the first demonstration of species-level laterality in manipulatory skills outside humans.

We can detect laterality in the manufacture of stepped tools (Fig. 2a) by New Caledonian crows (*Corvus moneduloides*) because this differs according to whether the left or the right leaf-edge is used¹. Birds carry out a well-defined sequence of precise cutting and ripping actions on both edges, with bill action being away from the tree trunk: they must therefore work from either right to left (left edges) or left to right (right edges). The right side of the head is generally towards the leaf edge in left-edge usage, and the left side in right-edge usage. A tool's finished shape is recorded in its unmistakable 'counterpart', the outline remaining on the leaf edge (Fig. 2a).

We collected 3,727 stepped-tool counterparts from 733 pandanus trees at 19 sites scattered throughout mainland Grande Terre, New Caledonia, the only island where crows

are known to make these tools. The sites were mostly at least 10 km apart and distributed along 300 km of the narrow, 400-km-long island at altitudes of 160–1,600 m above sea level. This sampling covered the geographical range of pandanus-toolmaking (G. R. H., personal observation).

The crow's use of left or right leaf-edges depends in part on the direction in which the leaves spiral. Clockwise-spiralling leaves provide easier access to left edges, and anticlockwise-spiralling leaves provide easier access to right edges. This access effect was overridden, however, by an island-wide preference for manufacturing tools from left edges (Fig. 2b).

Population-level footedness in manipulating food while feeding is known in parrots⁴. In great apes, population-level handedness has been reported in several manipulative activities in wild gorillas⁵ (*Gorilla g. beringei*) and in captive chimpanzees (*Pan troglodytes*)^{6–8}, but has not been shown so far in chimpanzees in the wild^{9,10}. Although there have been no systematic comparisons, it has been suggested that the level of complexity of a motor task and the associated neural-processing demands may correlate with the degree of lateralization^{8,11}.

The rather simple nature of chimpanzees' tool-making contrasts with the greater complexity of crows' stepped-tool manufacture¹. The shape of stepped tools is highly regular and mostly determined by the crows themselves, rather than by the raw material, imply-



Figure 1 New Caledonian crow. These birds make tools to probe for insects, including the 'crochet' tool illustrated.

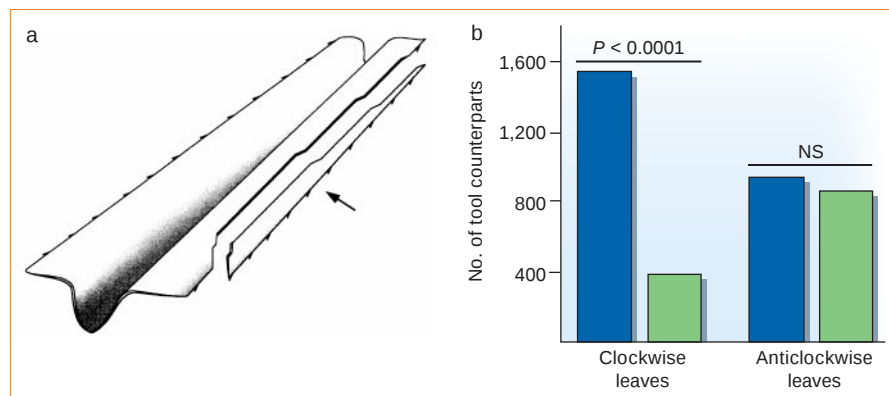


Figure 2 Lateralization in the manufacture of stepped tools from *Pandanus* sp. leaves by New Caledonian crows. a, A stepped tool (arrow), together with its counterpart shape on the left edge of a section of pandanus leaf (about 5 cm wide). Pandanus leaf rips easily along the strong, longitudinally parallel fibres, but must be cut laterally across the fibres. The tool is removed from right to left, beginning at the pointed end. Stepped-tool counterparts averaged 18.5 ± 6.30 cm long (mean $\pm$ s.d.m., $n = 3,418$) and 0.84 ± 0.24 cm wide ($n = 3,128$). b, The number of tool counterparts on the left (blue bars) and right (green bars) edges of leaves for each leaf-spiral direction. A 2×2 contingency test on the data (edge $\times$ spiral direction) found a significant access effect due to leaf-spiral direction (G-test of independence: $G_1 = 330.3$, $P < 0.0001$, $n = 3,727$), but there still exists a highly significant preference ($\chi^2_1 = 430.3$, $P < 0.0001$, $n = 3,727$) for manufacturing tools on left edges ($n = 2,463$) rather than right edges ($n = 1,264$). As individual birds are likely to have contributed more than one counterpart, this sample size is inflated. A conservative alternative is to treat each site as the unit of analysis. A Wilcoxon signed-ranks, matched-pairs test showed that the left-edged bias was still significant across the 19 sites ($T = 44$, two-tailed $P < 0.05$, $n = 19$). Sample sizes for tool counterparts varied between sites (196.2 ± 120.8 , $n = 3,727$), but the number of trees sampled with clockwise ($n = 369$) and anticlockwise ($n = 364$) leaves were almost identical (38.6 ± 24.4 trees per site, $n = 733$).

ing that a neural programme is involved in their manufacture¹. Laterality in crows may therefore reflect the complexity of stepped-toolmaking and a specialization of the right-eye/left-hemisphere system for non-spatial, sequential actions, as previously demonstrated in birds and mammals⁸.

It has been proposed that right-handedness in humans may be a consequence of the evolution of language, which is also predominantly left-hemispheric³. Our results favour the more general possibility that species-level lateralization is an adaptation for the efficient neural programming of complex sequential processing^{8,12}, of which language and right-handedness in humans and stepped-tool manufacture in crows are examples.

Gavin R. Hunt, Michael C. Corballis, Russell D. Gray

Department of Psychology, University of Auckland, Auckland 92019, New Zealand
e-mail: grhunt10@hotmail.com

- Hunt, G. R. *Proc. R. Soc. Lond. B* **267**, 403–413 (2000).
- Hunt, G. R. *Nature* **379**, 249–251 (1996).
- Corballis, M. C. *The Lopsided Ape: The Evolution of the Generative Mind* (Oxford University Press, Oxford, 1991).
- Harris, L. J. *Can. J. Psychol.* **43**, 369–396 (1989).
- Byrne, R. W. & Byrne, J. M. E. *Cortex* **27**, 521–546 (1991).
- Hopkins, W. D. *Psychonom. Bull. Rev.* **3**, 449–457 (1996).
- Hopkins, W. D. & Leavens, D. A. *J. Comp. Psychol.* **112**, 95–99 (1998).
- Bradshaw, J. L. & Rogers, L. J. *The Evolution of Lateral Asymmetries, Language, Tool Use, and Intellect* (Academic, San Diego, 1993).
- McGrew, W. C. & Marchant, L. F. *Int. J. Primatol.* **18**, 787–810 (1997).
- McGrew, W. C. & Marchant, L. F. *Behaviour* **138**, 329–358 (2001).
- Fagot, J. & Vauclair, J. *Psychol. Bull.* **109**, 76–89 (1991).
- Vallortigara, G., Rogers, L. J. & Bisazza, A. *Brain Res. Rev.* **30**, 164–175 (1999).

Light emission

A temperature-tunable random laser

Random lasers have fascinating emission properties that lie somewhere between those of a conventional laser and a common light-bulb. We have created a random laser that can be brought above and below its threshold for laser emission by small changes in its temperature, thereby creating a light source with a temperature-tunable colour spectrum. As a single random laser can be made as small as a grain of tens of micrometres in diameter, we expect our device to find application in photonics, temperature-sensitive displays and screens, and in remote temperature sensing.

Lasers are now commonplace — for example, they are used in industry and in hospitals, in bar-code scanners and compact-disc players. Conventional lasers are based on an optically active material and some sort of laser cavity that traps light for long enough for laser action to occur. A new type of laser source, known as a random laser¹, has been discovered that does not require a regular cavity but instead depends on a diffusive material such as a fine powder^{2–4}. In a random laser, light waves are trapped by multiple light scattering (that is, light diffusion⁵), which takes over the role of the cavity in a regular laser (Fig. 1). The emission of a random-laser source has a well defined colour spectrum and can be pulsed, just like a regular laser^{6–11}, although its emission is in several directions because of the intrinsic randomness of the system.

To create a random laser, light diffusion must be combined with light amplification, for instance by grinding a laser crystal into a fine powder or by adding diffusely scattering particles such as glass powder to a laser-dye solution. For our random-laser material, we infiltrated sintered glass with laser dye dissolved in a liquid crystal (see Fig. 1 legend). Random-laser action was evident as a strong narrowing of the emission spectrum (by up to a factor of 5) and an abrupt increase of the emitted intensity above an excitation energy of about 1 millijoule per pulse, as is commonly seen for random lasers based on laser dye.

The threshold at which a random laser starts to manifest is determined by the careful balance between gain and loss, as for a common laser. The losses for a random laser are due to light that escapes through the sample surface; the overall gain depends on the excitation energy and on how strongly the material scatters light (expressed in terms of a photon-diffusion constant). The more the material scatters light (the smaller the diffusion constant),

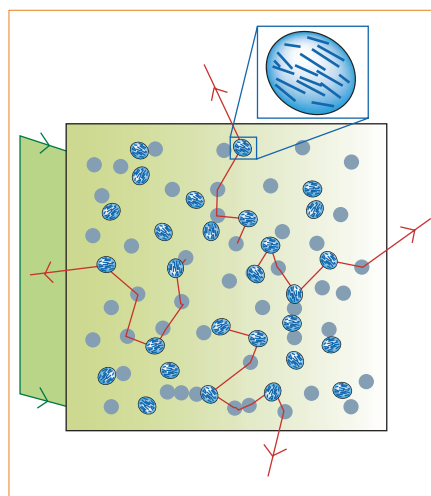


Figure 1 A temperature-tunable random laser. Light waves are multiply scattered by the sintered glass and liquid crystal and are amplified by the laser dye, which is excited by an external laser (green). The multiple scattering keeps the light inside the system for long enough for the amplification to become effective; when gain becomes larger than loss, the system will exhibit random-laser action. Our samples were made from SK11 glass (Schott) powders (pressed into discs of diameter 1 cm and thickness 1.3 mm). The resulting percolating porous structure was infiltrated at room temperature with the laser dye DCM (Lambdachrome 6500) dissolved in liquid crystal 4-cyano-4'-n-heptylbiphenyl (7CB) at various concentrations.

the longer the light is kept inside and the larger the overall gain can grow.

By using a liquid crystal inside our disordered material, we can have external control over the diffusion constant¹². Liquid crystals go through partially ordered phases when heated — each liquid-crystal phase has a different refractive index and therefore different scattering properties. When a random sample is infiltrated with a liquid crystal, its diffusion constant becomes strongly temperature-dependent, in much the same way that a polymer-dispersed liquid-crystal display changes its opacity with temperature¹³. For our samples, we observe a diffusion constant that increases gradually with temperature, with a more abrupt increase around the nematic–isotropic phase-transition temperature of the liquid crystal (in the nematic phase, the liquid-crystal molecules are aligned along a common axis, causing this phase to be birefringent; at higher temperatures, in the isotropic phase, the liquid crystal behaves as a normal isotropic liquid). This enables us to bring the random laser above and below its threshold by temperature-tuning of the diffusion constant.

The marked effect of different temperatures on the emission spectrum is shown in Fig. 2a. At low temperatures, there is random laser action (narrow spectrum), whereas at higher temperatures the random laser is below threshold (broad emission spectrum). By varying the temperature, the spectral width of this random laser

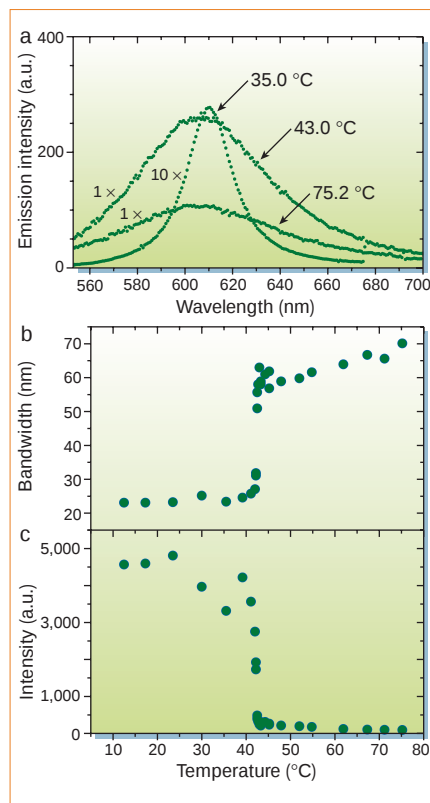


Figure 2 Action of a temperature-tunable random laser. The random-laser emission is characterized by exciting the sample with a frequency-double Nd:YAG laser and recording the spectrum of the diffuse emission (beam diameter, 1.9 mm; pulse duration, 14 ns; repetition rate, 10 Hz). **a**, Temperature dependence of the emission spectrum. At the highest temperature, the liquid crystal is in the isotropic phase and the laser is below its threshold (height at 35 °C is scaled by a factor of 10). **b**, Temperature dependence of the emission bandwidth. The bandwidth drops abruptly below 42.5 °C owing to the onset of laser action. **c**, Change in maximal emission intensity (arbitrary units, a.u.), showing an abrupt increase below the threshold temperature. By choosing different liquid-crystal/glass combinations the random laser can be designed with different tuning curves.

can be controlled (Fig. 2b); the abrupt reduction in bandwidth below 42.5 °C is due to the onset of laser action. The transition temperature corresponds, within experimental error, to the nematic–isotropic phase-transition temperature (42.8 °C) of the liquid crystal 7CB used here.

The onset of laser emission is nicely illustrated if we plot the observed intensity at the wavelength at which the emission spectrum is maximal (Fig. 2c). There is an abrupt increase in intensity below 42.5 °C; intensity then saturates at lower temperatures. This strong increase is possible due to repumping of the laser dye above threshold⁷. By choosing different tuning curves for the diffusion constant through selection of the sintered glass/liquid-crystal combination, the tunable random laser can be designed to show different spectral behaviour — for example, the temperature at which laser action occurs can be chosen and the abruptness or smoothness of the tuning

curves varied. This laser may find application as a source in active displays and temperature-sensitive screens and — because the threshold behaviour of tunable random lasers can be set to specific temperatures — in remote temperature sensing, for example in investigations of biological processes.

Diederik S. Wiersma*, Stefano Cavalieri†

*European Laboratory for Nonlinear Spectroscopy and Istituto Nazionale per la Fisica della Materia, Largo E. Fermi 2, 50125 Florence, Italy
e-mail: wiersma@lens.unifi.it

†European Laboratory for Nonlinear Spectroscopy, Dipartimento di Fisica, Università di Firenze, and Istituto Nazionale per la Fisica della Materia, Largo E. Fermi 2, 50125 Florence, Italy

1. Letokhov, V. S. Zh. Eksp. Teor. Fiz. **53**, 1442–1452 (1967) [Sov. Phys. J. Exp. Theoret. Phys. **26**, 835–840 (1968)].
2. Gouedard, C., Husson, D., Sauteret, C., Auzel, F. & Migus, A. J. Opt. Soc. Am. B **10**, 2358–2363 (1993).
3. Lawandy, N. M., Balachandran, R. M., Gomes, A. S. L. & Sauvin, E. Nature **368**, 436–438 (1994).
4. Wiersma, D. S. & Legendijk, A. Phys. Rev. E **54**, 4256–4265 (1996).
5. Sheng, P. Introduction to Wave Scattering, Localization, and Mesoscopic Phenomena (Academic, San Diego, 1995).
6. Zyuzin, A. Phys. Rev. E **51**, 5274–5278 (1995).
7. John, S. & Pang, G. Phys. Rev. A **54**, 3642–3652 (1996).
8. Berger, G. A., Kempe, M. & Genack, A. Z. Phys. Rev. E **56**, 6118–6122 (1997).
9. Beenakker, C. W. J. Phys. Rev. Lett. **81**, 1829–1832 (1998).
10. Cao, H. et al. Phys. Rev. Lett. **82**, 2278–2281 (1999).
11. Xunyu, J. & Soukoulis, C. M. Phys. Rev. Lett. **85**, 70–73 (2000).
12. Wiersma, D. S., Colocci, M., Righini, R. & Aliev, F. Phys. Rev. B **64**, 144208–1–144208–6 (2001).
13. Drzaic, P. S. Liquid Crystal Dispersions (World Scientific, Singapore, 1995).

COMMUNICATIONS ARISING

Seed development

Early paternal gene activity in *Arabidopsis*

Both parental genomes are expressed during embryogenesis, although the time of activation of the paternally inherited genes varies between organisms^{1,2}. Results reported by Vielle-Calzada *et al.*³ have suggested that delayed activation of the paternal genome seems to be the rule in plant development. We find, however, that during early embryogenesis in *Arabidopsis*, paternal genes are expressed and are sufficient for normal development. Our findings indicate that there is no overall maternal control of early embryogenesis, and that the contribution of the parental alleles needs to be assessed for each gene individually.

Arabidopsis embryos develop from a small egg cell within an ovule that consists largely of maternal tissue, making it difficult to determine the parental origin of embryonic transcripts by polymerase chain reaction with reverse transcription. *In situ* hybridization detects differential gene expression between the two daughter cells of the zygote⁴ without revealing the parental origin of the endogenous messenger RNA.

By contrast, uniparental gene activity can be monitored as transgene expression in embryos from crosses^{3,5}. Whereas paternal contribution implies post-fertilization expression, zygotic activity of the maternal allele can only be inferred when there is no evidence for pre-fertilization expression.

The *AtRPS5A::GUS* transgene strongly expresses β -glucuronidase (GUS) from the promoter of the gene encoding the small-ribosomal-subunit protein 5A in early embryogenesis. Fertilization of wild-type egg cells with transgenic pollen results in detectable GUS expression in most embryos as early as the two-cell stage (Fig. 1a), indicating that there is early activity of a paternally derived transgene.

Mutations in several *Arabidopsis* genes affect early embryogenesis, indicating that these genes function at the early stages of embryo development^{6–8}. Vielle-Calzada *et al.*³ propose that early embryogenesis depends on the maternal genome and that early defects due to a non-functional maternal allele cannot be complemented by a functional paternal allele. As subtle early-embryo phenotypes are difficult to classify unambiguously as either wild-type or mutant, we used a double-mutant genotype, *knolle keule*, in which cytokinesis is blocked from fertilization onwards, resulting in single-celled embryos with multiple nuclei⁷ (Fig. 1b). We found that

selfing of *knolle keule* heterozygous plants gave the expected one-quarter of single-celled embryos, whereas almost all embryos from wild-type pollination were normal (Fig. 1c). We conclude that functional paternal alleles are sufficient for normal embryogenesis from the earliest stage.

By using a two-component expression system⁹ in *Arabidopsis*, we found additional evidence for early expression of paternally inherited transgenes (results not shown). However, expression from the maternal genome was consistently stronger than from the paternal genome, as shown here by reciprocal crossing between a driver line with an epidermis-specific gene promoter¹⁰ fused to a chimaeric transcription factor, and a line with the corresponding artificial response element controlling the GUS reporter gene (Fig. 1d). These differences in expression between maternal and paternal alleles may occur in early (data not shown) as well as in later embryogenesis.

Our findings seem to contradict the idea of overall maternal control of early embryogenesis³, but may be reconciled with those of Vielle-Calzada *et al.*³ if paternal gene expression is attenuated rather than silenced, or if it is affected by locus-specific rather than genome-wide imprinting. Whether or not the activity of a specific paternal gene is sufficient during early plant embryogenesis might depend on

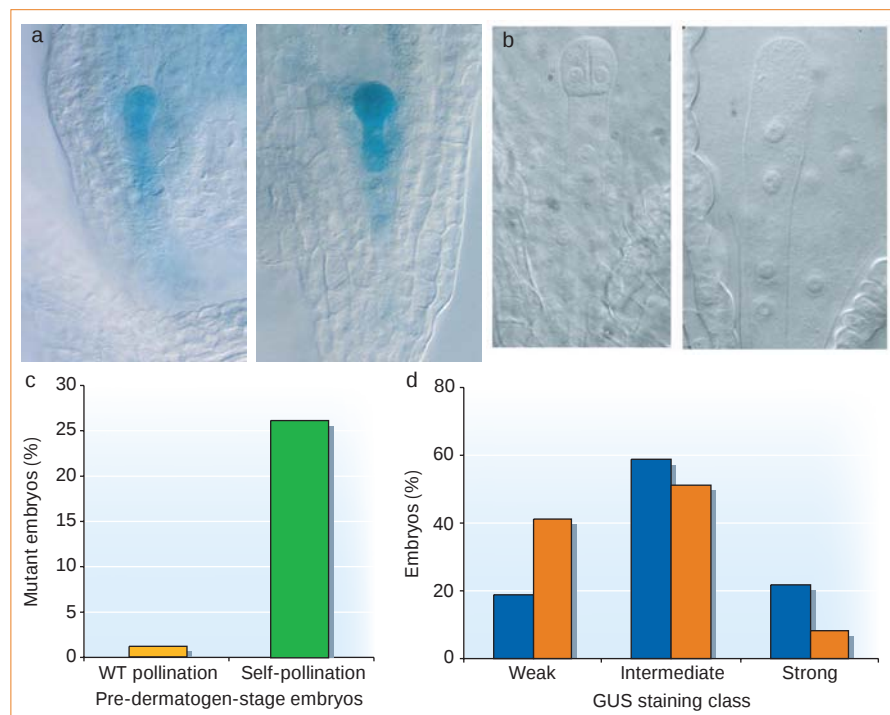


Figure 1 Paternal gene activity during early embryogenesis. **a**, β -Glucuronidase (GUS) activity in embryos from a cross between an *AtRPS5A::GUS* male and a wild-type female. **b**, Wild-type two-cell-stage embryo (left) and *knolle keule* double mutant (right). **c**, Selfing of a heterozygous *knolle keule* double mutant *cis*-line leads to 26.2% mutant embryos ($n = 172$) versus 1.2% mutant embryos ($n = 173$) from wild-type (WT) pollination. Embryos from one-cell to eight-cell stages were counted. **d**, GUS activity in protoderm cells of heart to torpedo-stage embryos from reciprocal crosses between a *pLTP1::GAL4-VP16* line and a *pUAS::GUS* reporter line. Blue bars ($n = 144$), maternally expressed, and orange bars ($n = 132$), paternally expressed *pLTP1::GAL4-VP16*, indicating the proportion of embryos classified as weak, intermediate or strong expressors.

differential gametic imprinting, as well as on the amount of gene product needed for biological function.

Dolf Weijers*, Niko Geldner†, Remko Offringa*, Gerd Jürgens†

**Institute of Molecular Plant Sciences, Leiden University, 2333 AL, Leiden, The Netherlands.*

†Zentrum für Molekularbiologie der Pflanzen, Universität Tübingen, 72076 Tübingen, Germany

e-mail: gerd.juergens@zmbp.uni-tuebingen.de

1. Surani, M. A. *et al. Development* (suppl.) 89–98 (1990).
2. Lehmann, R. & Nüsslein-Volhard, C. *Dev. Biol.* **119**, 402–417 (1987).
3. Vielle-Calzada, J.-P., Baskar, R. & Grossniklaus, U. *Nature* **404**, 91–94 (2000).
4. Lu, P., Porat, R., Nadeau, J. A. & O'Neill, S. D. *Plant Cell* **8**, 2155–2168 (1996).
5. Springer, P. S., Holding, D. R., Groover, A., Yordan, C. & Martienssen, R. A. *Development* **127**, 1815–1822 (2000).
6. Mayer, U., Büttner, G. & Jürgens, G. *Development* **117**, 149–162 (1993).
7. Waizenegger, I. *et al. Curr. Biol.* **10**, 1371–1374 (2000).
8. Hamann, T., Mayer, U. & Jürgens, G. *Development* **126**, 1387–1395 (1999).
9. <http://www.plantsci.cam.ac.uk/Haseloff/GAL4/mGAL4.html>
10. Thoma, S. *et al. Plant Physiol.* **105**, 35–45 (1994).

Vielle-Calzada et al. reply — Our results, based on a study of 20 loci, indicate that the contributions by the maternal and paternal genome to early seed development in *Arabidopsis* are not equivalent, as evidenced by a lack of detectable paternal gene activity during the first few divisions after fertilization. As these loci are distributed throughout the genome, we inferred that early embryo and endosperm development are mainly under maternal control, but this may not be true for every locus and, as in X-chromosome inactivation¹, we would expect some loci to escape this silencing mechanism. We did not claim that maternal control is complete, but suggested that the activity of many genes during early embryo and endosperm formation could depend solely on transcription of the maternally inherited allele before and/or after fertilization.

Previously, early seed formation was thought to involve transcription from both parental copies immediately following fertilization, and maternal effects were considered rare or non-existent². The time at which paternal activity can first be detected, however, is likely to vary from embryo to embryo and from gene to gene in different nuclei, as in *Drosophila*³. Weijers *et al.* report paternal expression of *AtRPS5A::GUS* as early as the two-cell stage, confirming that transcription in the zygote is not the rule for paternally inherited alleles, whereas transcription from maternal alleles has been demonstrated immediately after fertilization of the central cell⁴. We do not know what percentage of embryos show early *AtRPS5A::GUS* expression, nor the relative paternal and maternal activity, but there may also be less pronounced parent-of-origin differences.

New evidence supporting the non-equivalence of maternal and paternal genomes during early seed development is based on experiments with reporter genes^{5–8} and genetic assays revealing maternal effects of genes thought to act purely zygotically⁶ (S. Gilmore and C. Somerville, personal communication; J. Moore and U. G., unpublished results). Whether and at what stage expression of the paternal allele is sufficient for normal development will depend on the level of activity required for gene function. In a two-component transactivation system, no paternal activity was found during early seed development using *pOp::GUS* reporter lines with several activator lines⁸. Some early defects were evident with a *pOp::BARNASE* reporter, however, suggesting that paternal transcription is very low but is sufficient to cause *BARNASE*-induced defects in some embryos⁸. These results confirm the non-equivalence of maternal and paternal contributions to early seed development. Like imprinted genes in mammals, this difference is probably not absolute and may be due to different levels of maternal and paternal transcripts.

Our titration experiments indicated a difference in transcript levels of at least 80-fold for genes we tested by PCR. Weijers *et al.* report an expression difference in reciprocal crosses with *UAS::GUS* at the heart to torpedo stage (Fig. 1d), when we showed that both parental alleles are active at other loci we tested; indeed, this differential expression translates into an absence of detectable paternal activity at earlier stages using the *pOp::GUS* reporter system⁸. For some genes, such as *KEULE* or *KNOLLE*, low paternal expression may be sufficient for normal development, although very early defects (such as developmental delay) that are rescued by a paternal wild-type allele may be difficult to detect by scoring multinucleate

embryos. Moreover, rescue of an early embryonic phenotype by a paternal wild-type allele provides no evidence against differences in parental transcript levels.

Although the exact time of paternal activation was not central to our report, most evidence so far suggests that no consistent paternal gene activity can be detected in the embryo or endosperm for several cell divisions. The results of Weijers *et al.* do not contradict our findings, but instead represent possible exceptions to a general rule. Specific genes that are important during early development (for example, those involved in cytokinesis that are distinctly regulated in the female gametophyte and the zygote⁹) may be under selection for earlier expression and be specifically activated early in development. Further investigation is required into how common early-expressing paternal genes are, and how maternal and paternal expression differs quantitatively.

Jean-Philippe Vielle-Calzada*, Ramamurthy Baskar†, Ueli Grossniklaus†

**Cinvestav-Plant Biotechnology Unit, Irapuato, Mexico; †Institute of Plant Biology, University of Zürich, 8008 Zürich, Switzerland*

e-mail: grossnik@botinst.unizh.ch

1. Goto, T. & Monk, M. *Microbiol. Mol. Biol. Rev.* **62**, 362–368 (1998).
2. Howell, S. H. *Molecular Genetics of Plant Development* (Cambridge Univ. Press, Cambridge, 1999).
3. Pritchard, D. K. & Schubiger, G. *Genes Dev.* **10**, 1131–1142 (1996).
4. Vielle-Calzada, J.-P. *et al. Genes Dev.* **13**, 2971–2982 (1999).
5. Luo, M., Bilodeau, P., Dennis, E. S., Peacock, W. J. & Chaudhury, A. M. *Proc. Natl Acad. Sci. USA* **97**, 10637–10642 (2000).
6. Springer, P. S., Holding, D. R., Groover, A., Yordan, C. & Martienssen, R. A. *Development* **127**, 1815–1822 (2000).
7. Sorensen, M. B., Chaudhury, A. M., Robert, H., Bacharel, E. & Berger, F. *Curr. Biol.* **11**, 277–281 (2001).
8. Baroux, B., Blanvillain, R. & Gallois, P. *FEBS Lett.* **509**, 11–16 (2001).
9. Mayer, U., Herzog, U., Berger, F., Inzé, D. & Jürgens, G. *Eur. J. Cell Biol.* **78**, 100–108 (1999).

corrections

Night-time predation by Steller sea lions

G. L. Thomas, R. E. Thorne

Nature **411**, 1013 (2001)

We stated that our acoustic surveys in Prince William Sound since 1993 and infrared surveys since 2000 suggested that these sea lions “feed exclusively” on herring. However, it has been drawn to our attention that this statement is misleading. In clarification, the sea lions were selectively targeting the relatively shallow (0–50-m depth) schools of Pacific herring (*Clupea pallasii*) at night as a source of winter forage to the exclusion of relatively larger and deeper (150–250 m) concentrations of walleye pollock.

Transatlantic robot-assisted telesurgery

J. Marescaux, J. Leroy, M. Gagner, F. Rubino, D. Mutter, M. Vix, S. E. Butner, M. K. Smith

Nature **413**, 379–380 (2001)

The correct address of the third author of this communication is Division of Laparoscopic Surgery at Mount Sinai School of Medicine and Mount Sinai Medical Centre, New York 10029, USA.

Peptide antibiotics in mast cells of fish

Umaporn Silphaduang, Edward J. Noga

Nature **414**, 268–269 (2001)

The concentrations listed in Table 1 are in $\mu\text{g ml}^{-1}$.

erratum

Nitrate flux in the Mississippi River

G. F. McIsaac, M. B. David, G. Z. Gertner, D. A. Goolsby

Nature **414**, 166–167 (2001).

In Fig. 1 of this communication, the line referred to as “black” is in fact blue; also, in the fourth line of the third column, *P* should be greater than 0.05.

Breakdown of Fermi-liquid theory in a copper-oxide superconductor

R. W. Hill*, Cyril Proust^{†‡}, Louis Taillefer*, P. Fournier^{†‡} & R. L. Greene[†]

* Canadian Institute for Advanced Research, Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada

† Center for Superconductivity Research, Department of Physics, University of Maryland, College Park, Maryland 20742, USA

The behaviour of electrons in solids is well described by Landau's Fermi-liquid theory, which predicts that although electrons in a metal interact, they can still be treated as well defined fermions, which are called 'quasiparticles'. At low temperatures, the ability of quasiparticles to transport heat is given strictly by their ability to transport charge, as described by a universal relation known as the Wiedemann–Franz law, which hitherto no material has been known to violate. High-temperature superconductors have long been thought to fall outside the realm of Fermi-liquid theory, as suggested by several anomalous properties, but this has yet to be shown conclusively. Here we report an experimental test of the Wiedemann–Franz law in the normal state of a copper-oxide superconductor, (Pr,Ce)₂CuO₄, which reveals that the elementary excitations that carry heat in this material are not fermions. This is compelling evidence for the breakdown of Fermi-liquid theory in high-temperature superconductors.

Landau's Fermi-liquid theory is the definitive theory of electrons in metals¹, or more generally fermions in condensed matter, and is a major landmark of twentieth-century physics. For example, it is the necessary foundation for the theory of superconductivity by Bardeen, Cooper and Schrieffer (BCS)². In essence, it says that even in the presence of interactions the low-energy excitations of a system of mobile fermions can still be described in terms of well defined fermionic particles, called 'quasiparticles', with charge e , spin 1/2 and a mass m^* , the latter being renormalized by interactions. The Wiedemann–Franz (WF) law is one of the basic properties of a Fermi liquid, reflecting the fact that the ability of a quasiparticle to transport heat is the same as its ability to transport charge, provided it cannot lose energy through collisions. Reported³⁰ as an empirical observation by Wiedemann and Franz in 1853, the law states that the heat conductivity κ and the electrical conductivity σ of a metal are related by a universal constant:

$$\frac{\kappa}{\sigma T} = \frac{\pi^2}{3} \left(\frac{k_B}{e} \right)^2 \equiv L_0 \quad (1)$$

where T is the absolute temperature, k_B is Boltzmann's constant and $L_0 = 2.45 \times 10^{-8} \text{ W } \Omega \text{ K}^{-2}$ is Sommerfeld's value³ for the Lorenz ratio, $L \equiv \kappa/\sigma T$.

The linear power of temperature in equation (1) comes from the linear temperature dependence of the fermionic specific heat, through the relation between heat transport and heat capacity. In kinetic theory,

$$\kappa = \frac{1}{3} c v l \quad (2)$$

where c is the specific heat of the carriers, v is their average velocity and l their mean free path. Provided the mean free path between collisions is independent of temperature, as it is for the strictly elastic processes that operate at $T \rightarrow 0$, the heat conduction has the same T dependence as the specific heat. So for electrons, it is also linear in T , and given by equation (1). By contrast, boson-like excitations such as phonons or magnons give rise to a low-temperature specific heat and a heat conductivity as $T \rightarrow 0$ which are both proportional to T^3 .

Theoretically, electrons are predicted to obey the WF law at $T \rightarrow 0$ in a very wide range of environments⁴: in both three dimensions or two dimensions (but not strictly in one dimension), for any strength of disorder and interaction (in the scaling theory of disorder and interaction)⁵, scattering⁶ and magnetic field⁷. Experimentally, the WF law does appear to be universal at $T \rightarrow 0$: no materials have been reported to violate it. (We note that a departure from equation (1) has in fact been reported for silver⁸, but is in contradiction with other measurements; refs 9, 10, and R.W.H. and E. Boaknin, unpublished results). This is true not only of simple metals like copper^{9,10}, but also of systems with strong electron correlations—such as the heavy fermion compounds CeAl₃ (ref. 11), CeCu₆ (ref. 12) and UPt₃ (ref. 13)—or with highly anisotropic conduction—such as the quasi-two-dimensional system Sr₂RuO₄ (ref. 14) (a ruthenate which is isostructural to the cuprate La₂CuO₄), the quasi-one-dimensional organic conductor (TMTSF)₂ClO₄ (ref. 15) or the two-dimensional electron gas in a MOSFET¹⁶. Even in the presence of so-called non-Fermi-liquid behaviour, where the electrical resistivity is seen to deviate from the standard T^2 dependence, the WF law still holds at $T \rightarrow 0$, as observed in CeNi₂Ge₂ (ref. 17). All this strongly suggests that the ground state of every metal investigated thus far is a Fermi liquid. Of course, if the metallic state gives way to a superconducting state at low temperature, then the WF law is entirely violated, as the charge is no longer transported by fermions but by Cooper pairs, which carry no entropy.

In 1986, high-temperature superconductivity was discovered in a class of copper oxides, and fifteen years hence the fundamental question about these materials still remains: are they host to a new state of matter? In particular, are they Fermi liquids in their ground state, once superconductivity is removed (either by application of a magnetic field or by doping)? Anderson¹⁸ proposed early on that the copper oxides are indeed fundamentally different from other metals in that they are Mott insulators. When a small number of electrons or holes are doped into the CuO₂ planes of their crystal structure, these materials can conduct charge reasonably well and also form a superconducting state. It is believed that the basic excitations of a doped Mott insulator are not fermionic quasiparticles as in other metals—with charge, spin and heat all carried by one and the same particle. Instead, the electron is thought to 'fractionalize' into a neutral spin-1/2 fermion called a 'spinon' and a spinless charge- e boson called a 'holon', or a 'chargon', a phenomenon usually called 'spin-charge' separation¹⁹.

‡ Present addresses: Département de Physique, Université de Sherbrooke, Sherbrooke, Québec J1K 2R1, Canada (P.F.); Laboratoire National des Champs Magnétiques Pulsés, 143 Avenue de Rangueil, BP 4245, 31432 Toulouse, Cedex 04, France (C.P.).

Even though this bold proposal for a fundamentally new state of matter continues to be at the heart of several current theories of the copper oxides (ref. 20 and references therein), there has been little experimental evidence for such spin–charge separation. On the contrary, for example, a BCS/Fermi-liquid theory of *d*-wave quasiparticles appears to work rather well in the superconducting state of hole-doped copper oxides, at least at low energy and near optimal doping²¹. Conceptually, a simple way to investigate the possibility of spin–charge separation is to measure the properties of charge transport and compare them to those of either spin or entropy transport.

Here we report on the first test of this kind, where both the charge conductivity σ and the heat conductivity κ of a copper oxide superconductor are measured in the normal state at $T \rightarrow 0$. We observe a good, metallic-like nearly temperature-independent charge transport at low temperature without the corresponding heat transport expected for the WF law. The normalized Lorenz number L/L_0 goes from being much less than unity at $T \rightarrow 0$ to being greater than unity above 0.2 K. This strongly suggests spin–charge separation and is a compelling demonstration that the ground state of copper oxides can fall outside the realm of Fermi-liquid theory.

The material

The copper oxide material used for this study is $\text{Pr}_{2-x}\text{Ce}_x\text{CuO}_{4-y}$ (PCCO), with a doping concentration of $x = 0.15$, which is roughly the optimum value for maximizing the transition temperature for superconductivity, where $T_c \approx 20$ K. The addition of x Ce atoms on Pr sites adds x electrons to the CuO_2 planes of the parent insulating compound, the Mott insulator Pr_2CuO_4 . PCCO is the electron-doped analogue of the hole-doped material $\text{La}_{2-x}\text{Sr}_x\text{CuO}_{4-y}$ (LSCO), wherein the addition of x Sr atoms on La sites removes x electrons from (adds x holes to) the CuO_2 planes. In LSCO, the optimum doping is also at approximately $x = 0.15$, with a maximum $T_c \approx 40$ K. For reasons that are not clear, both materials have critical temperatures that are significantly lower than that of several other copper oxides, such as $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ (Y-123), $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ (Bi-2212) and $\text{Tl}_2\text{Ba}_2\text{CuO}_{6+\delta}$ (Tl-2201), which all have a maximum T_c around 90 K. Nevertheless, all copper oxides share the same basic crystal structure, namely that of a stack of CuO_2 planes, and the same generic dependence on doping. The technical advantage of LSCO and PCCO is that the magnetic field needed to destroy superconductivity is correspondingly lower, with upper critical field values of ~ 50 T and ~ 10 T, respectively. Since these fields can be achieved in the laboratory, it is possible to investigate their low-temperature transport properties in the normal state, such as was done for charge conductivity on LSCO²² and on PCCO²³.

Crystals

The single crystal of PCCO used in this study was grown with a flux technique described elsewhere^{24,25}, using an Al_2O_3 crucible and an oxygen reduction treatment at 1,000 °C. The concentration of Ce was fixed at $x = 0.15$. Contacts were made with silver epoxy, diffused at 500 °C for 1 hour, and were used to measure both electrical resistivity and thermal conductivity. The electrical resistance of contacts was typically $\sim 1 \Omega$. The crystal was in the shape of a platelet, 37 μm thick and 720 μm wide, with 1.0 mm separation between voltage contacts. The superconducting transition temperature T_c was obtained from three different measurements: at the end of the drop in resistivity, at the end of the diamagnetic drop in susceptibility and at the onset of a small peak in thermal conductivity. In all cases, $T_c = 20$ K to within less than ± 0.5 K. The width of the transition is 6 K in resistivity, 3 K in magnetization and about 1 K in thermal conductivity. The broad resistive transition, seen both as a function of temperature and magnetic field, is most probably due to an inhomogeneous distribution of oxygen near the

surface. The much narrower transition obtained in thermal conductivity shows that this inhomogeneity is confined to a negligible volume of the sample, and therefore does not affect bulk measurements such as heat transport. We note that all results presented in this paper were reproduced on several other single crystals grown in similar conditions, with comparable T_c and resistivity.

Charge transport

The temperature dependence of the electrical resistivity, ρ , is shown in Fig. 1 for different magnetic fields applied parallel to the *c* axis of the tetragonal crystal structure. The current is made to flow in the CuO_2 planes, where the conduction is some 10^4 times better than perpendicular to the planes (at room temperature). Above 25 K, the temperature dependence $\rho(T)$ is not linear as in most optimally doped copper oxides, but follows a power law $a + bT^\alpha$ with $\alpha \approx 1.7$, and has a magnitude characteristic of a reasonably good metal: $a \approx 25 \mu\Omega \text{ cm}$ and $\rho(300 \text{ K}) \approx 140 \mu\Omega \text{ cm}$. The effect of a magnetic field is rapidly to suppress the superconducting transition initially, but then more slowly at high fields, giving rise to a curve $T_c(H)$ that has strong positive curvature. From the magnetoresistance at 0.3 K shown in Fig. 1, lower inset, we can see that the entire sample has completed its transition to the normal state by 13.5 T. However, the bulk of the sample is no longer superconducting above 8 T, the field beyond which the low-temperature heat transport ceases to evolve (see Fig. 2).

The overall behaviour of the resistivity in PCCO is similar to that of hole-doped copper oxides when overdoped. Indeed, when doped to have a T_c of 15 K, Tl-2201 has a resistivity which grows as $T^{1.75}$ with values of $10 \mu\Omega \text{ cm}$ at low temperature and $180 \mu\Omega \text{ cm}$ at room temperature, and a resistive critical field with a strong upward curvature reaching a value of 16 T at $T \rightarrow 0$ (ref. 26). One difference is the slight upturn seen at low temperature in PCCO and not in Tl-2201. The resistivity in the normal state (at 14 T) increases by about 10% in going from 8 K down to 0.3 K (see Fig. 1). Similar but much larger upturns were observed in $\text{Nd}_{2-x}\text{Ce}_x\text{CuO}_4$ (NCCO)^{27,28} and PCCO²⁵ at lower carrier concentrations; these were attributed to weak localization. In support of this interpretation is the suppres-

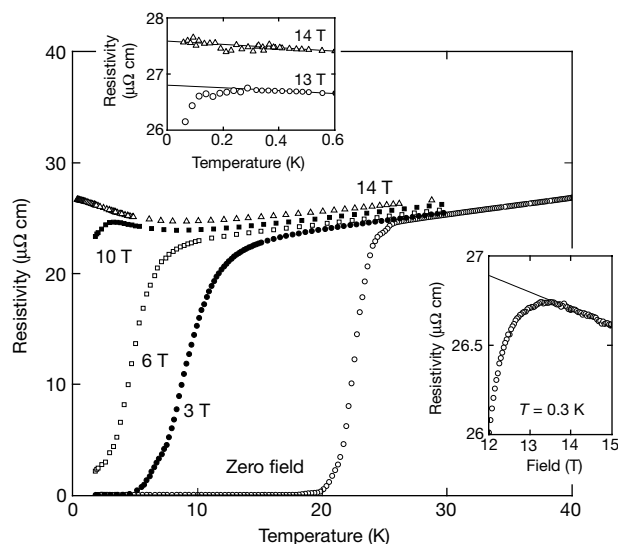


Figure 1 Electrical resistivity of PCCO versus temperature for a current in the basal plane at different values of magnetic field applied normal to the plane. Upper inset, zoom of low-temperature dependence at 13 T and 14 T, the latter being shifted upwards by $0.8 \mu\Omega \text{ cm}$ for clarity. Lines are linear fits to the data below 1.0 K. Lower inset, resistivity versus magnetic field at $T = 0.3$ K, showing that the crystal has fully completed its transition to the normal state by 13.5 T. The line is a linear fit to the negative magnetoresistance between 13.5 and 15 T.

sion of the upturn by a transverse magnetic field, also observed in our samples (see lower inset of Fig. 1). This negative magnetoresistance eventually flattens to a constant beyond 25 T, equal to $25.2 \mu\Omega \text{ cm}$ at 0.5 K (R.W.H., unpublished results).

Thus, once the superconducting order has been completely suppressed throughout the sample, namely in a magnetic field larger than 13.5 T, the charge transport of this copper-oxide material at low temperature is that of a fairly good quasi-two-dimensional metal, with a residual resistivity $\rho(T \rightarrow 0) = \rho_0 = 26.8 \mu\Omega \text{ cm}$ in 14 T, which corresponds to a conductivity per CuO_2 plane of $60 e^2/h$ (that is, $k_F l \approx 60$).

Heat transport

Fifteen years after the discovery of high-temperature superconductivity, very little is known about the normal state of copper oxides at low temperature. The ability to suppress superconductivity in optimally doped PCCO with a field of only 14 T provides a rare opportunity to shed light on the fundamental nature of the electron system in these materials. We did this by measuring the transport of heat as $T \rightarrow 0$. The thermal conductivity $\kappa(T)$ of PCCO is shown in Fig. 2 at temperatures below 0.2 K, for heat flowing parallel to the CuO_2 planes. The data is plotted as κ/T versus T^2 in order easily to separate the contributions of phonons and electrons, using the fact that phonons are expected to have a heat conductivity which varies as T^3 when their mean free path reaches the size of the crystal at sufficiently low temperature. A straight line extrapolating to zero in κ/T versus T^2 is indeed the asymptotic ($T \rightarrow 0$) behaviour observed in insulators²⁹, including the Mott insulator $\text{YBa}_2\text{Cu}_3\text{O}_6$ (that is, Y-123 without any doped carriers)³⁰. This is also the behaviour of conventional (*s*-wave) superconductors. The absence of a residual linear term κ_0/T , defined as the value of κ/T as $T \rightarrow 0$, means that there are no mobile fermionic excitations at low temperature. In insulators, this is owing to the lack of electronic carriers, while in superconductors it is because electrons have all formed Cooper pairs at $T \ll T_c$. The heat is then conducted entirely by phonons. In contrast, a residual linear term is observed when fermionic excitations are present, as in the following three instances: (1) in metals, with a residual linear term given precisely by the WF law; (2) in the vortex state of *s*-wave superconductors, where a magnetic field induces a residual linear term by generating delocalized

quasiparticles³¹; and (3) in *d*-wave superconductors, where zero-energy quasiparticles are induced by impurity scattering³².

Superconducting state

In the absence of a magnetic field, there is only a very small residual linear term in the thermal conductivity of PCCO. From data on four different crystals, we find $\kappa_0/T = 0.03\text{--}0.07 \text{ mW K}^{-2} \text{ cm}^{-1}$. This is considerably smaller than the residual linear term observed in three hole-doped copper oxides, at optimal doping: $\kappa_0/T = 0.14, 0.15$ and $0.11 \text{ mW K}^{-2} \text{ cm}^{-1}$, in Y-123 (refs 30, 33), Bi-2212 (refs 34, 35) and LSCO³⁶, respectively. Within BCS theory applied to a *d*-wave superconductor, this fermionic residual heat conduction is expected, arising from zero-energy quasiparticles induced by impurity scattering near the nodes in the $d_{x^2-y^2}$ gap function³². In the case of Bi-2212, the excellent quantitative agreement between theory and experiment has been viewed as a strong validation of Fermi-liquid quasiparticle theory for the superconducting state^{21,34}, at least at optimal doping and low energies.

Several tentative explanations may be offered for the weakness of a residual linear term in PCCO. The first possibility is the localization of *d*-wave quasiparticles (ref. 37 and references therein). It was invoked recently as a possible cause for the absence of a residual linear term in $\text{YBa}_2\text{Cu}_4\text{O}_8$ (Y-124)³⁸. However, it is not clear why localization should occur in Y-124 and not in Y-123, given that these two copper oxides have comparable structure and properties. It seems more plausible in PCCO, because of the highly two-dimensional nature of its conduction—some 100 times more anisotropic than either Y-124 or LSCO. On the other hand, Bi-2212 is equally anisotropic, yet shows delocalized quasiparticles. A second possibility is the absence of nodes in the gap structure of PCCO. This would eliminate any residual linear term (at least in zero magnetic field). The symmetry of the order parameter need not be *s*-wave, but could have the form $d + ix$, where *x* is a subdominant component of either *s* or d_{xy} symmetry. The possibility of a $T = 0$ transition in the order parameter from *d* to $d + ix$ as a function of doping has been raised in the context of a possible quantum critical point in the phase diagram of copper oxides³⁹. For PCCO, however, this would seem to be in conflict with evidence for a rather pure $d_{x^2-y^2}$ symmetry, as determined in a recent tricrystal experiment⁴⁰. The other difficulty is that, as we shall see below, a magnetic field normal to the CuO_2 planes never induces a residual linear term, implying that the minimum gap would have to be a sizeable fraction of the maximum gap (that is, *x* comparable to *d*). This would have to be reconciled with the penetration depth data^{41,42}. A third possibility⁴³, is the removal of the nodes as a result of static stripe ordering.

Normal state

As seen in Fig. 2, the basic effect of a magnetic field is rapidly to increase the thermal conductivity of PCCO (at finite temperature). Given that the phonon conductivity cannot exceed its value at zero field (where the phonon mean free path has reached its maximum value), this increase must be due to excitations of electronic origin. The field-induced conduction (over and above the zero-field curve) grows rapidly, and approximately linearly, with magnetic field at low fields, and eventually saturates at high fields, where κ is absolutely constant above 8 T. This is compelling evidence that the thermodynamic upper critical field in this sample is at most 8 T and the bulk of the crystal is in the normal state beyond that. Indeed, in all known superconductors, whether *s*-wave or *d*-wave, the electronic thermal conductivity is always affected by the magnetic field when in the vortex state and seen to saturate above H_{c2} (see for example ref. 44). Such an H_{c2} agrees reasonably well with the criterion of zero resistance, achieved at about 6 T, and also agrees with the H_{c2} values reported previously in thin films of PCCO^{23,45} and NCCO^{45,46} at $x = 0.15$. A roughly linear growth of κ with magnetic field is also what is seen in the borocarbide superconductor $\text{LuNi}_2\text{B}_2\text{C}$ (ref. 31), but with an important difference: in

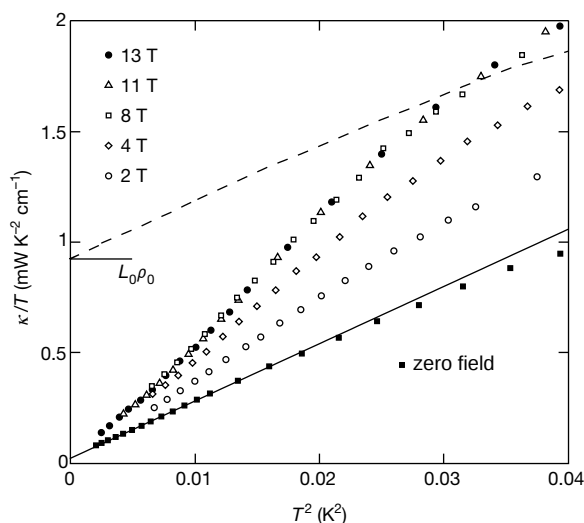


Figure 2 Thermal conductivity of PCCO for a heat current in the basal plane, plotted as κ/T versus T^2 , at different values of the magnetic field applied normal to the plane. The solid line is a linear fit to the zero-field data below 130 mK. The dashed line shows the behaviour of a Fermi liquid with the residual resistivity ρ_0 of this sample, calculated as the sum of a constant electronic linear term via equation (1) and a phonon conductivity given by the solid line (zero-field data).

$\text{LuNi}_2\text{B}_2\text{C}$, the field induces a linear term at $T \rightarrow 0$. Furthermore, this finite residual linear term ultimately reaches a value in the normal state given precisely by the WF law. In striking contrast, a magnetic field, even as high as 13 T, almost twice H_{c2} , never induces a residual linear term in PCCO. Given the known value of the electrical resistivity at $T \rightarrow 0$ and 13 T, namely $\rho_0 = 26.8 \mu\Omega \text{ cm}$, the WF law predicts a residual linear term $\kappa_0/T = L_0/\rho_0 = 0.91 \text{ mW K}^{-2} \text{ cm}^{-1}$, as drawn in Fig. 2. (Weak localization is not expected to alter the WF law appreciably⁵, as verified experimentally in layered graphene⁴⁷.)

This complete violation of the WF law shows that the non-superconducting ground state of this copper-oxide material is not a Fermi liquid. The charge carriers responsible for the good electrical conduction do not display the expected fermionic heat transport.

It is instructive to consider the effect of increasing the temperature above zero. Contributions to the heat transport from the electron system (κ_e) and from the ionic lattice (κ_{ph}) will add to give the measured thermal conductivity: $\kappa(T) = \kappa_e(T) + \kappa_{ph}(T)$. These two systems are coupled. It is clear that phonons are scattered by electrons, since the crystal conducts heat better at zero field than at 13 T beyond 0.6 K. In other words, as the temperature is raised, inelastic scattering processes become increasingly effective at limiting the phonon heat flow, and we expect the dominant process to be electron–phonon scattering.

Given this observed electron–phonon coupling, the total phonon scattering in field is the sum of whatever scatters phonons in zero field plus the added scattering that comes from the field-induced electronic excitations. The thermal resistivity of phonons due to electron scattering is given by $W_e^{ph} = BT^{-2}$, where the constant B depends on the strength of the electron–phonon coupling, the electronic density of states, and various other factors⁴⁸. This means that the phonon conductivity at low T is given by $\kappa_{ph} = \kappa(0)[1 + \kappa(0)W_e^{ph}]^{-1}$, where $\kappa(0)$ is the measured zero-field conductivity, assumed to be essentially all phononic. We can put a reasonable lower bound on the electron–phonon coupling parameter B by requiring that the resulting κ_e/T does not decrease with T . Adjusting B such that κ_e/T is constant above 0.3 K gives the curve shown in Fig. 3 for $H = 13 \text{ T}$ (circles). The same can be done for other fields with the resulting values of $\kappa_e/T = 0.6, 1.2$ and $1.9 \text{ mW K}^{-2} \text{ cm}^{-1}$ and corresponding values of $B = 0.22, 0.4$, and $0.68 \text{ K}^3 \text{ m W}^{-1}$ for $H = 2, 4$ and 8 (or 13) T, respectively. (In comparison, electron–phonon coupling in copper gives $B = 4\text{--}8 \text{ K}^3 \text{ m W}^{-1}$; ref. 48). We note the self-consistency of this analysis whereby B scales with the value of κ_e/T (above 0.3 K); in other words, by increasing the field from 2 to 4 T—thereby increasing the density of electronic excitations—one increases both the electronic heat conduction and the electron scattering of phonons.

From Fig. 3 we see that $\kappa_e(T)/T$ exceeds the value of L_0/ρ that is appropriate for a Fermi liquid. In other words, the low-energy excitations of the electron system violate the WF law not only at $T \rightarrow 0$, where $\kappa_e(T)/T \ll L_0/\rho_0$, but also at finite temperatures, where $\kappa_e(T)/T > L_0/\rho$ (above 0.2 K). It is important to appreciate that qualitatively the latter violation is independent of our choice of B , in the sense that even if B is set to zero, κ_e/T still exceeds L_0/ρ , in this case by 30% at 0.25 K. We stress that $B = 0$ is unphysical, as it would lead to a negative κ_e/T for $T > 0.6 \text{ K}$. It is also independent of our assumption that $\kappa(0)$ is entirely phononic; indeed, if some part of $\kappa(0)$ is in fact electronic, then the violation turns out to be even more pronounced. In other words, while there is no accurate way to extract the electronic contribution at 13 T, the fact that the WF law is violated is independent of our model for the phonon conduction and the electron–phonon scattering. The result $\kappa_e(T)/T \approx 2L_0/\rho_0$ above 0.3 K is obtained in several crystals, even though $\rho(T_c)$ varies in the range $20\text{--}40 \mu\Omega \text{ cm}$.

We have investigated the possibility that our measurement gradually ceases to detect the electronic heat current as electrons fall out of thermal equilibrium with the phonon bath when the

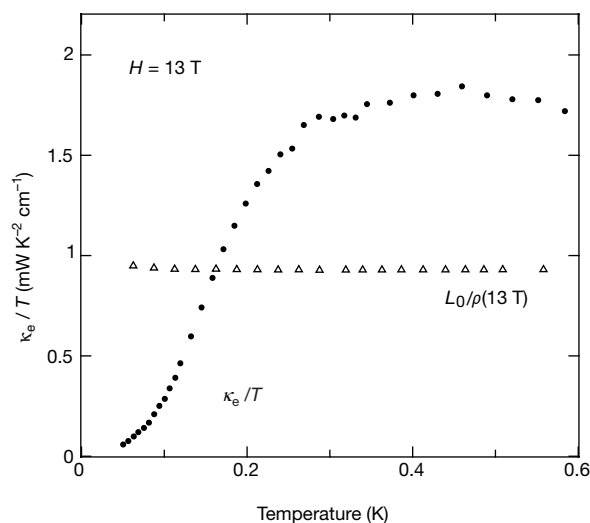


Figure 3 Comparison of charge conductivity $\sigma(T) = 1/\rho(T)$, plotted as $L_0/\rho(T)$ (triangles), and electronic heat conductivity κ_e , plotted as κ_e/T (circles), as a function of temperature in the normal state at $H = 13 \text{ T}$. The electronic contribution to the heat conduction is the difference between the measured $\kappa(13 \text{ T})$ and the phonon contribution $\kappa_{ph}(13 \text{ T})$, estimated in the text. In a Fermi liquid, the curve of κ_e/T would lie precisely on the data for L_0/ρ_0 . Below $T = 0.15 \text{ K}$, $\kappa_e \approx T^{3.6}$.

temperature approaches $T = 0 \text{ K}$. We can imagine this happening when the electrical contact resistance between the heater and the sample is too large and the heat is predominantly carried by the phonons. This might then explain the drop in κ_e/T measured below 0.3 K or so (see Fig. 3). Two facts lead us to rule out electron–phonon decoupling as the main mechanism for the drop. First, identical data is obtained by sending heat directly through the electron system using photons⁴⁹. Secondly, a similar drop is observed in samples of optimally doped $(\text{La, Sr})_2\text{CuO}_4$, where we have achieved electrical contact resistances that are 100 times lower than in our PCCO samples, namely $\sim 10 \text{ m}\Omega$. We therefore believe this low-temperature drop is intrinsic and not specific to PCCO.

Discussion

The copper oxide PCCO is the first material, to our knowledge, to violate the Wiedemann–Franz law. The breakdown of this robust signature of Fermi-liquid theory suggests that the fundamental entities that carry heat, charge (and spin) in the copper oxides are not the usual Landau quasiparticles. The fact that the electron system conducts heat in a way that is largely unrelated to the way it carries charge at low temperature points to the existence of neutral excitations responsible for much of κ_e . If these neutral excitations are fermions, that is, spinons, they should contribute a constant term to κ_e/T , as may well be the case for $T > 0.3 \text{ K}$. The drop seen at the very lowest temperatures then remains to be explained, and the possibility of spinon localization should be considered.

This study is limited to a single carrier concentration, near optimal doping, in the whole phase diagram. It will be interesting to investigate the evolution with doping of this emerging picture. If the WF law is satisfied in overdoped copper oxides, where Fermi-liquid theory is generally assumed to hold, our results on optimally doped PCCO suggest that a violation of the law is a property of underdoped copper oxides. A comprehensive study will shed light on the gradual breakdown of Fermi-liquid theory in the copper oxides as the carrier concentration is reduced towards the Mott insulating state. □

Received 10 August; accepted 27 September 2001.

- Landau, L. D. The theory of a Fermi liquid. *Sov. Phys. JETP* **3**, 920–925 (1957).
- Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Theory of superconductivity. *Phys. Rev.* **108**, 1175–1204 (1957).

3. Ashcroft, N. W. & Mermin, N. D. *Solid State Physics* 322 (HRW, Philadelphia, 1976).
4. Kearney, M. J. & Butcher, P. N. Thermal transport in disordered systems. *J. Phys. C. Solid State Phys.* **21**, L265–L270 (1988).
5. Castellani, C., DiCastro, C., Kotliar, G., Lee, P. A. & Strinati, G. Thermal conductivity in disordered interacting-electron systems. *Phys. Rev. Lett.* **59**, 477–480 (1987).
6. Chester, G. V. & Tellung, A. The law of Wiedemann and Franz. *Proc. Phys. Soc.* **77**, 1005–1013 (1961).
7. Smrcka, L. & Streda, P. Transport coefficients in strong magnetic fields. *J. Phys. C. Solid State Phys.* **10**, 2153–2161 (1977).
8. Gloos, K., Mitschka, C., Pobell, F. & Smeibidl, P. Thermal conductivity of normal and superconducting metals. *Cryogenics* **30**, 14–18 (1990).
9. Anderson, A. C., Peterson, R. E. & Robichaux, J. E. Thermal and electrical conductivity of Ag and Pt below 1 K. *Phys. Rev. Lett.* **20**, 459–461 (1968).
10. Rumbo, E. R. Transport properties of very pure copper and silver below 8.5 K. *J. Phys. F Metal Phys.* **6**, 85–98 (1976).
11. Ott, H. R., Marti, O. & Hulliger, F. Low temperature thermal conductivity of CeAl₃. *Solid State Commun.* **49**, 1129–1131 (1984).
12. Amato, A. *et al.* Transport properties of CeCu₆ at very low temperature. *J. Magn. Magn. Mater.* **63–64**, 300–302 (1987).
13. Suderow, H., Brison, J. P., Huxley, A. & Flouquet, J. Thermal conductivity and gap structure of the superconducting phases of UPT₃. *J. Low Temp. Phys.* **108**, 11–30 (1997).
14. Tanatar, M. A. *et al.* Thermal conductivity of superconducting Sr₂RuO₄ in oriented magnetic fields. *Phys. Rev. B* **63**, 064505-1–064505-7 (2001).
15. Belin, S. & Behnia, K. Thermal conductivity of superconducting (TMTSF)₂ClO₄: evidence for a nodeless gap. *Phys. Rev. Lett.* **79**, 2125–2128 (1999).
16. Syme, R. T., Kelly, M. J. & Pepper, M. Direct measurement of the thermal conductivity of a two-dimensional electron gas. *J. Phys. Condens. Matter.* **1**, 3375–3380 (1989).
17. Kambe, S. *et al.* Spin-fluctuation mediated thermal conductivity around the magnetic instability of CeNi₂Ge₂. *J. Low Temp. Phys.* **117**, 101–112 (1999).
18. Anderson, P. W. The resonating valence bond state in La₂CuO₄ and superconductivity. *Science* **235**, 1196–1198 (1987).
19. Kivelson, S. A., Rokhsar, D. S. & Sethna, J. P. Topology of the resonating valence bond state: Solitons and high-*T_c* superconductivity. *Phys. Rev. B* **35**, 8865–8868 (1987).
20. Senthil, T. & Fisher, M. P. A. Z₂ gauge theory of electron fractionalization in strongly correlated systems. *Phys. Rev. B* **62**, 7850–7881 (2000).
21. Orenstein, J. & Millis, A. J. Advances in the physics of high-temperature superconductivity. *Science* **288**, 468–474 (2000).
22. Boebinger, G. S. *et al.* Insulator-to-metal crossover in the normal state of La_{2-x}Sr_xCuO₄ near optimum doping. *Phys. Rev. Lett.* **77**, 5417–5420 (1996).
23. Fournier, P. *et al.* Insulator-metal crossover near optimal doping in Pr_{2-x}Ce_xCuO₄: Anomalous normal-state low temperature resistivity. *Phys. Rev. Lett.* **81**, 4720–4723 (1998).
24. Peng, J. L., Li, Z. Y. & Greene, R. L. Growth and characterization of high-quality single crystals of R_{2-x}Ce_xCuO_{4-y} (R=Nd,Sm). *Physica C* **177**, 79–85 (1991).
25. Brinkmann, M. *et al.* Crystal growth of high-*T_c* superconductors Pr_{2-x}Ce_xCuO_{4-δ} with substitutions of Ni and Co for Cu. *J. Crystal Growth* **163**, 369–376 (1996).
26. Mackenzie, A. P. *et al.* Resistive upper critical field of Ti₂Ba₂CuO₆ at low temperatures and high magnetic fields. *Phys. Rev. Lett.* **71**, 1238–1241 (1993).
27. Harus, G. I. *et al.* Two-dimensional weak localisation effects in high temperature superconductor Nd_{2-x}Ce_xCuO_{4-δ}. *Sov. JETP* **116**, 1–12 (1999).
28. Fournier, P. *et al.* Anomalous saturation of the phase coherence length in underdoped Pr_{2-x}Ce_xCuO₄. *Phys. Rev. B* **62**, R11993–R11996 (2000).
29. Thacher, P. D. Effect of boundaries and isotopes on the thermal conductivity of LiF. *Phys. Rev.* **156**, 975–988 (1967).
30. Taillefer, L. *et al.* Universal heat conduction in YBa₂Cu₃O_{6.9}. *Phys. Rev. Lett.* **79**, 483–486 (1997).
31. Boaknin, E. *et al.* Highly anisotropic gap function in the borocarbide superconductor LuNi₂B₂C. *Phys. Rev. Lett.* (in the press); preprint cond-mat/0012436 at (<http://xxx.lanl.gov>) (2001).
32. Durst, A. C. & Lee, P. A. Impurity-induced quasiparticle transport and universal-limit Wiedemann-Franz violation in *d*-wave superconductors. *Phys. Rev. B* **62**, 1270–1290 (2000).
33. Chiao, M. *et al.* Quasiparticle transport in the vortex state of YBa₂Cu₃O_{6.9}. *Phys. Rev. Lett.* **82**, 2943–2946 (1999).
34. Chiao, M. *et al.* Low-energy quasiparticles in cuprate superconductors: A quantitative analysis. *Phys. Rev. B* **62**, 3554–3558 (2000).
35. Behnia, K. *et al.* Features of heat conduction in organic and cuprate superconductors. *J. Low-Temp. Phys.* **117**, 1089–1098 (1999).
36. Taillefer, L. & Hill, R. W. Heat transport in high-temperature superconductors. *Phys. Can.* **56**, 237–240 (2000).
37. Senthil, T. *et al.* Quasiparticle transport and localisation in high *T_c* superconductors. *Phys. Rev. Lett.* **81**, 4704–4707 (1998).
38. Hussey, N. E. *et al.* Absence of residual quasiparticle conductivity in the underdoped cuprate YBa₂Cu₃O₈. *Phys. Rev. Lett.* **85**, 4140–4143 (2000).
39. Vojta, M., Zhang, Y. & Sachdev, S. Quantum phase transitions in *d*-wave superconductors. *Phys. Rev. Lett.* **85**, 4940–4943 (2000).
40. Tsuei, C. C. & Kirtley, J. R. Phase-sensitive evidence for *d*-wave pairing symmetry in electron-doped cuprate superconductors. *Phys. Rev. Lett.* **85**, 182–185 (2000).
41. Kokales, J. D. *et al.* Microwave electrodynamics of electron-doped cuprate superconductors. *Phys. Rev. Lett.* **85**, 3696–3699 (2000).
42. Prozorov, R. *et al.* Evidence for nodal quasiparticles in electron-doped cuprates from penetration depth measurements. *Phys. Rev. Lett.* **85**, 3700–3702 (2000).
43. Granath, M. *et al.* Nodal quasiparticles and coexisting orders in striped superconductors. *Phys. Rev. Lett.* **87**, 167011-1–167011-4 (2001).
44. Belin, S. *et al.* Probing the upper critical field of κ-(BEDT-TTF)₂Cu(NCS)₂. *J. Superconduct.* **12**, 497–500 (1999).
45. Kleefisch, S. *et al.* Possible pseudogap behaviour of electron-doped high-temperature superconductors. *Phys. Rev. B* **63**, 100507-1–100507-4 (2001).
46. Osofsky, M. S. *et al.* in *Proc. 10th Anniversary HTS Workshop* 284–287 (World Scientific, Singapore, 1996).
47. Bayot, V. *et al.* Evidence for weak localisation in the thermal conductivity of a quasi-two-dimensional electron system. *Phys. Rev. Lett.* **65**, 2579–2582 (1990).
48. Berman, R. *Thermal Conduction in Solids* 145 (Clarendon, Oxford, 1976).
49. Gutsmedl, P., Probst, C. & Andres, K. Low temperature calorimetry using an optical heating method. *Cryogenics* **31**, 54–57 (1991).
50. Wiedemann, G. & Franz, R. Ueber die Wärme-Leitungsfähigkeit der Metalle. *Ann. Phys.* **89**, 497–532 (1853).

Acknowledgements

We thank C. Lupien, E. Boaknin, M. Sutherland, M. Chiao, R. Gagnon, J. Brooks, L. Balicas and B. Brandt for their invaluable help in various aspects of the measurements. This work was supported by the Canadian Institute for Advanced Research and funded by NSERC of Canada. L.T. acknowledges the support of a Premier's Research Excellence Award from the Government of Ontario. The work in Maryland was supported by the NSF Division of Condensed Matter Physics.

Correspondence and requests for materials should be addressed to L.T. (e-mail: Louis.Taillefer@utoronto.ca).

Travelling waves and spatial hierarchies in measles epidemics

B. T. Grenfell*, O. N. Bjørnstad*† & J. Kappey*

* Zoology Department, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

† 501 ASI Building, Departments of Entomology and Biology, Penn State University, University Park, Pennsylvania 16802, USA

Spatio-temporal travelling waves are striking manifestations of predator–prey and host–parasite dynamics. However, few systems are well enough documented both to detect repeated waves and to explain their interaction with spatio-temporal variations in population structure and demography. Here, we demonstrate recurrent epidemic travelling waves in an exhaustive spatio-temporal data set for measles in England and Wales. We use wavelet phase analysis, which allows for dynamical non-stationarity—a complication in interpreting spatio-temporal patterns in these and many other ecological time series. In the pre-vaccination era, conspicuous hierarchical waves of infection moved regionally from large cities to small towns; the introduction of measles vaccination restricted but did not eliminate this hierarchical contagion. A mechanistic stochastic model suggests a dynamical explanation for the waves—spread via infective ‘sparks’ from large ‘core’ cities to smaller ‘satellite’ towns. Thus, the spatial hierarchy of host population structure is a prerequisite for these infection waves.

Travelling waves, arising essentially from activator–inhibitor dynamics^{1–3}, are predicted by theory in a range of host–natural enemy systems^{1,4–8}. However, except as the product of invasion dynamics⁹, empirical observations of waves are comparatively rare—especially repeated periodic waves associated with host–natural enemy population cycles^{7,8,10–12}. Even where waves are dynamically possible, they may not be detected because of a lack of spatio-temporal data at the appropriate resolution. More subtly, spatial heterogeneities in population density or demography^{8,13} and temporal changes in parameters (resulting, for example, from vaccination against disease^{1,6}), can significantly alter the detection and dynamics of spatio-temporal waves. The integration of models and data to explore these interactions between spatial dynamics and host demography is often prevented by a lack of demographic information.

Childhood microparasitic infections—in particular, historical measles epidemics in developed countries—provide sufficiently detailed spatio-temporal data on disease incidence and host demography^{14,15} to address these issues. The task is aided by epidemiological models^{16–21}, which capture both the nonlinear dynamics of childhood epidemics as a function of local population size¹⁶ and the impact of significant environmental forcing. This forcing mainly comprises seasonality in transmission, due to schooling patterns¹⁷, and longer-term variations in susceptible recruitment, due to birth-rate variations and the onset of vaccination^{20,22}. Such long-term temporal changes in dynamic rates give rise to non-stationarity in the resulting ecological time series.

The detection of temporal and spatio-temporal oscillations in time series is greatly complicated by non-stationary temporal variations in dynamical behaviour (such as changes in mean, variance, period of oscillations, and so on). In particular, trends or sudden jumps in cycle period complicate the search for temporal and spatial patterns, since conventional frequency-domain analyses assume stationarity^{23,24}. Figure 1a illustrates the periodic but non-stationary dynamics of weekly measles notifications for London for the period 1944–94. There are marked changes in epidemic period between the 1940s and 1950s, as well as into the vaccine era (after 1968).

We apply wavelet time series analysis^{23–26} to describe the non-stationarity in the period of recurrent epidemics of measles in England and Wales. Wavelet phase angles also reveal rapid spatio-temporal waves of infection, originating from regional centres.

Finally we use a refined epidemiological model to suggest how such waves can arise in seasonal environments, and explore the role of spatial heterogeneities in creating them.

Local measles dynamics

Measles epidemics in developed countries generally exhibit seasonal cycles and longer-term (generally biennial) major epidemics^{18,19,27}. However, the relative importance of the seasonal versus multiannual cycles varies with time. These features of the dynamics are clearly seen from a wavelet spectral analysis of the time series of weekly measles notifications for London from 1944 to 1994 (Box 1 and Fig. 1). The local wavelet power spectrum (Fig. 1b) shows the importance of the different oscillatory periods as a function of time. The seasonal cycles and (in general) biennial major epidemics are obvious, as also is the long-term non-stationarity in the period of the major epidemic. The main long-term change in dynamics accompanies the onset of vaccination in 1968 (Fig. 1b). After this, the annual periodicity is less marked and the intervening major epidemics (now of lower amplitude) gradually increase in period, compared to the pre-vaccine era. This gradual transition coincides with a steady increase in vaccine uptake from around 50% in the 1970s to around 90% in the late 80s (Fig. 1d). Theory predicts^{18,20}, and previous time-series analyses have confirmed^{20,28}, that vaccination should generate an increase in the epidemic period. Here we use the temporal dimension of the wavelet analysis to reveal the progressive nature of this increase (Fig. 1b).

To generate a synoptic picture of transitions in measles dynamics across 354 administrative areas of England and Wales (see Methods), we focus on changes through time in the dominant epidemic period (Fig. 1d). The regional pattern is remarkably consistent with the detailed analysis for London: biennial epidemics before widespread vaccination are followed by epidemics of longer period through the vaccine era. A second important feature of the pre-vaccination era is the reduction in inter-epidemic interval (to under 2 years) coinciding with the 1945–47 and 1962–65 ‘baby booms’²². All these transitions in disease dynamics are driven by ‘extrinsic’ variations in recruitment rate of susceptibles²⁰ (Fig. 1d) through changes in birth rates (discounted by vaccine uptake in the vaccination era). The analysis includes hundreds of locations, ranging from large cities (where there are regular epidemics) to small towns (where disease dynamics are strongly influenced by stochasticity²⁷). Consequently, these results give an unusually

Box 1

Wavelet time series analysis

We introduce the basic principles of the method^{23–26}, as applied to measles in London. Traditional Fourier analysis partitions the total power (variance) of the time series between sinusoidal components at different frequencies²⁹. If we apply this approach to series with a marked change in cycle period through time (such as Fig. 1a), the resulting spectrum reflects both scales of variation, but tells us nothing about their sequence. This is particularly problematic if the data show unusual ‘spikes’ and interruptions, or in the presence of smooth changes in periodicity²³.

By contrast, the essence of the wavelet approach is locality in time as well as frequency. Rather than a sinusoid, the method is based on a wavelet function, which can explore local variations in frequency. The last decade has seen an explosion of applications of wavelets in mathematics and its applications²⁵; this has led to a wide range of wavelet functions.

In this paper, we use the Morlet wavelet function²³ (see Methods and Fig. 1e). The Morlet function is essentially a damped complex exponential, which can capture local (in time) cyclical fluctuations in the time series. As with all wavelets, the frequency–time range over which it does this is set by a scale parameter, s . In general, wavelet scale is related to the conventional Fourier period of oscillations (Methods).

A local wavelet power spectrum (LWPS; see Methods) of the weekly London measles time series (Fig. 1a) is presented in Fig. 1b–d. Power is colour-coded as shown. Seasonality in the epidemic generates a peak in power at approximately one year. The major peak in power varies around two years, as described in the text. The superimposed parabola is the cone of influence (see Methods), which measures the extent of edge effects. Only significant ($P < 0.05$) power is shown.

Time and scale averaging

A useful feature of the LWPS is that it can be averaged across both frequency and time (Methods²³). Figure 1c shows the global wavelet

spectrum, estimated by averaging the LWPS across time (the dotted line is the lower limit of significance). The global spectrum is analogous to the traditional Fourier spectrum. We note the annual and roughly biennial peaks of power. Other wavelet methods used here, such as the analysis of phase differences and smoothing by reconstruction of important frequencies, are discussed in the Methods.

As described in the text, Fig. 1d shows a synoptic summary across all spatial locations in England and Wales of the period of major epidemics (in the range 1.5–5 years). The blue line indicates the average of this period across 354 spatial locations in England and Wales; the thickness of the line represents ± 1 standard error of the mean, s.e.m. The red line shows vaccination (p) rate. The black line is the effective recruitment rate of susceptibles, $B(1 - p)$, where B is the birth rate (the relevant axis is suppressed: recruitment drops from a maximum of 0.021 per individual in 1947 to near zero by 1994). Figure 1d shows the progressive change from period 2 to period 3 dynamics over the 1970s and 80s. However, the subsequent increase to 4 years in the late 1980s and 90s should be interpreted cautiously, since it may reflect the secular decline at the end of the time series, caused by the major increase in vaccination uptake over this period.

Wavelet methods have considerable potential as tools for ecological time series and spatial analyses³⁸. However, like many ‘local’ statistical methods, they need a lot of data—features can be detected only at a given time and frequency if the underlying process is sufficiently well sampled. It is also important to try a variety of wavelet functions²⁴. A number of continuous and discrete functions can capture the basic patterns shown here; however, the Morlet function gives the clearest picture.

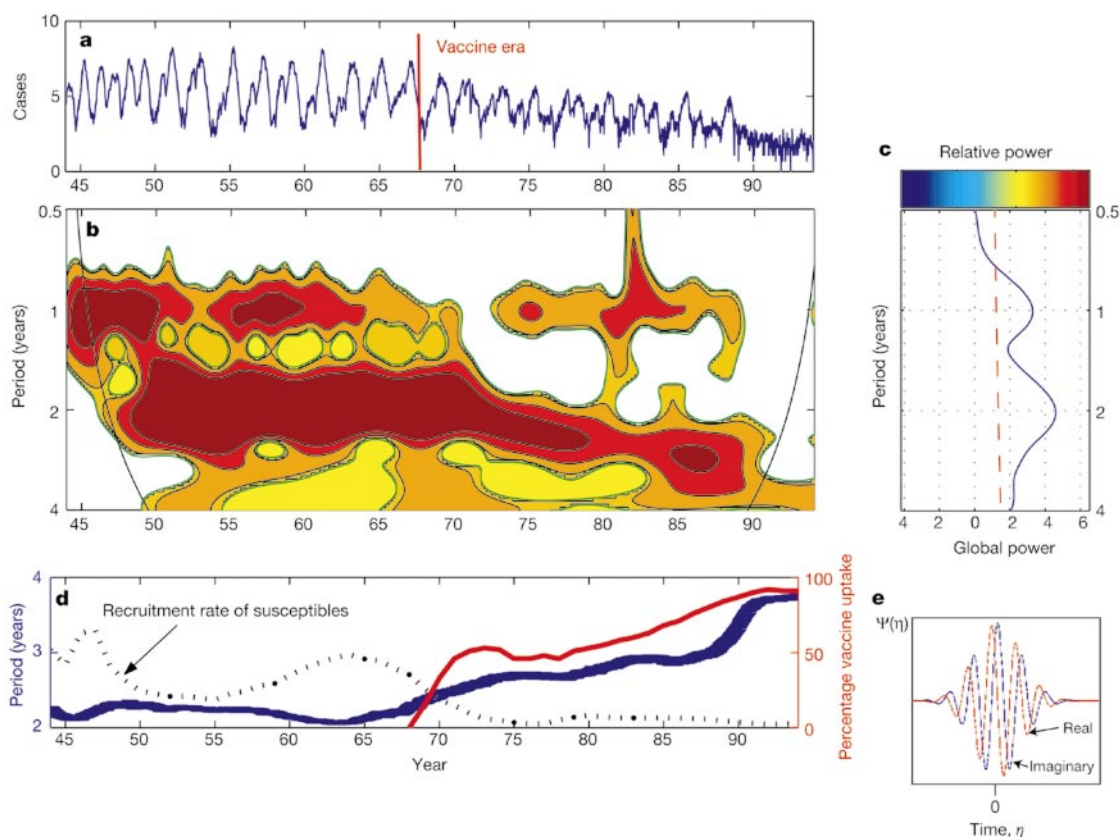


Figure 1 Wavelet time series analysis for the log-transformed weekly London measles time series (see Box 1 for details). **a**, The time series. **b**, Local wavelet power spectrum (LWPS); power is colour coded as shown on the key at top right. **c**, Global wavelet

spectrum. **d**, Summary of temporal changes in the dominant epidemic period, averaged across towns and cities in England and Wales. **e**, Shape of the Morlet wavelet used in the analysis.

detailed picture of the influence of changes in host demography on epidemic dynamics.

Travelling waves and wavelet phase angles

We investigate spatio-temporal patterns by calculating the phase difference between epidemics at different locations. For cycles with a given period, the wavelet analysis generates a phase angle at each time step (Methods)²³. This contrasts with standard multivariate spectral analysis, which can capture phase relationships at different periods²⁹, but does not allow a description of how they change with time. Because the raw phase spectrum is defined at each time step and period, it is difficult to interpret directly. We therefore focus on the phase of the relatively well defined (generally biennial) major epidemics, using wavelet reconstruction²³ (see Methods). We investigate the wavelet phases for periods between 1.5 and 3 years, using the wavelet reconstruction to calculate the average cycle and associated phase across these periods at each time step. This procedure allows naturally for observed spatio-temporal variations in major epidemic period and phase (Box 1).

The phase analysis is well illustrated by the pre-vaccination records for Norwich and Cambridge, because epidemics in these relatively close cities (90 km apart) were out of phase during the 1950s (ref. 30, Fig. 2). We also include results for London for comparison. The raw series (Fig. 2a) shows that major epidemics in Cambridge were aligned with London (and the national pattern) and occurred in odd years (1951, 1953, 1955, and so on). Norwich, by contrast, had epidemics that peaked in even years for the 15 years following World War Two. Figure 2c shows the major epidemic pattern for the three cities as revealed by the wavelet reconstruction. This confirms that the main epidemic period in London was predominantly biennial from around 1950; before that, epidemics tended to be more annual (Fig. 1a, b). More dramatically, the

analysis highlights the early phase difference between Cambridge and Norwich (Fig. 2c–e), followed by the change in Norwich's epidemic phase from odd to even years in the late 1950s (we return below to the cause of Norwich's unusual even-year major epidemic timing). The analysis indicates that the biennial cycles of Cambridge and Norwich were 100 to 160° (7–11 months) out of phase for 16 years (Fig. 2e), after which they locked more closely into phase. Figure 2e also shows that major epidemics in Cambridge lagged several weeks behind those in London.

We observe from Fig. 2 that the phase difference between major epidemics changes relatively smoothly through time. This advances our objective of interpreting the complex spatio-temporal pattern of measles epidemics in terms of the spatial pattern of phase differences. For instance, phase-locked fluctuations (for example, refs 7 and 3) should result in zero phase-difference across the map, whereas travelling waves^{1,4,5} should generate a phase difference that increases with distance.

We focus first on the 1950–66 period—after the 1947 baby boom but before vaccination—when the 2-year epidemic cycle was particularly pronounced (Fig. 1a). Figure 3a maps the average phase difference relative to London for the whole urban pre-vaccination data set (954 locations). Most, though not all, places lag behind London (88% have negative phase difference). Concentrating on the London region (Fig. 3c), there is a clear wave of infection moving away from the capital city. The wave is particularly well defined up to 30 km from London (Fig. 3c) with a wave speed of around 5 km per week. Figure 3d reveals the existence of a similar wave around Manchester. (We note that the focus of this wave is less well defined because there are several large cities—such as Leeds and Liverpool—that feed into the epidemic dynamics of North-western England.) The measles waves are illustrated dynamically in the Supplementary Information: the movie shows the raw incidence

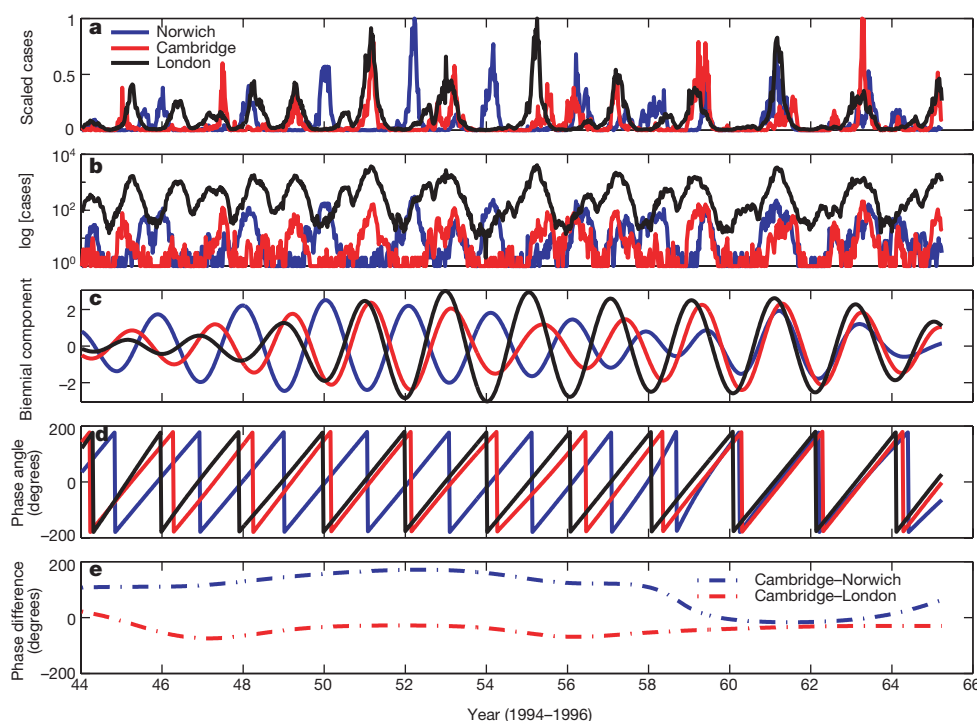


Figure 2 Wavelet phase analysis of weekly measles reports for Cambridge, Norwich and London in the pre-vaccine era. **a**, Pattern of weekly case reports. **b**, Logarithmic values of number of cases reported: $\log[x + 1]$. **c**, Major epidemic (mainly biennial) component of the three series, reconstructed from wavelet spectral analysis (Box 1); the series were reconstructed from components in the period range 1.5 to 3 years, to allow for a flexible period ('multiannual' is probably therefore a better term than biennial, though the

predominant variation is over 2 years). **d**, Average phase angles of the reconstructions in **c**. **e**, Phase difference between Cambridge and the other two cities, based on the phase angles in **d**; to remove spurious 'jumps in phase difference, the raw phase difference (θ) is constrained within $\pm 180^\circ$, by the transformation: $\text{mod}(\theta + 540, 360) - 180$, where $\text{mod}(x, y)$ is x modulo y .

and phase analysis for the whole of England and Wales across seven epidemic cycles.

The time-averaged phase map (Fig. 3a) highlights the hierarchical nature of the waves, which move from large population centres to the surrounding hinterland throughout the 1950s and 60s. During this era, several hierarchical waves have their foci in large cities and spread progressively to more distant small towns (see also ref. 14). The dominant waves are those associated with London (Fig. 3c) and the northwestern industrial centre (Fig. 3d), but the hierarchical spread of infection is a consistent feature of the phase map, as is reflected in its spatial correlation structure (Fig. 4a, b; see below). An interesting feature of the pattern is that, although London—the biggest city—leads the epidemics in surrounding areas (Fig. 3a, c), the major epidemic appears on average 4–6 weeks before London in the urban northwest (Fig. 3a, d), moving from Liverpool and Manchester into the north Midlands. We speculate that this may arise from the unusual annual dynamics of Liverpool (which are due to high local birth rates³⁰), effectively ‘subsidizing’ the growth of the epidemic in Manchester and Leeds and thereby sparking off early hierarchical waves in this region. Such regional heterogeneities are likely to be an interesting line of inquiry in future work. A second notable feature is the distinct pattern during the 1950s of even-year epidemics in the cities and villages of East Anglia, centred on Norwich (Fig. 3a, b). We speculate on the origin of this pattern in Box 2.

To quantify the geographic extent of the epidemic waves, we use spatial phase coherency functions (PCF)—calculated as the (non-parametric) spatial correlation function⁷ of the major epidemic phases (Fig. 4a; see Methods). These functions express the correlation in phase angles at different locations as a function of intervening distance. During the pre-vaccination era, phases in nearby

locations are generally found to be highly correlated. The correlation declines with distance to a distinct minimum at around 250 km, which possibly crudely reflects the spatial extent of the two dominant hierarchical waves. The PCF reveals that the local phase coherence was significantly higher than the regional average to a distance of around 100 km (Fig. 4a). In the vaccine era (Fig. 4a), the PCF resembles the pre-vaccination function in shape, but the average coherence is consistently and significantly lower. The extent of significant local phase coherence dropped to around 75 km during the 1970s. Extensions of the phase analysis (not shown) indicate that the hierarchical waves moving from large centres persist through the 1980s. However, the spatial extent of the waves dropped further (local coherence is only significant to 35 km), as high vaccination rates induced irregular epidemics later in the vaccine era³².

Phase coherence effectively measures the relative timing of epidemics. A complementary approach is to consider the synchrony⁷ of the epidemic time series (see Methods), which also reflects how their relative amplitudes covary. The pattern of synchrony in measles (Fig. 4b) follows the qualitative pattern of the phase coherency: declining with distance and with a lower mean in the vaccine era. However, the synchrony of epidemics is significantly less than their phase coherence, especially in the vaccine era. This indicates that vaccination induces stronger variations in the amplitude of epidemics than in their relative phase. Previous studies^{32–34} have documented reductions in synchrony of measles epidemics as a result of vaccination. The present analysis reveals the detailed architecture of the change (Fig. 4b).

A ‘forced forest fire’ model for measles waves

The measles waves depart from the assumptions of classical theoretical models predicting travelling waves in two important ways. First, local transmission is strongly seasonally forced; second, the spatial distribution of the host is inherently very heterogeneous. We

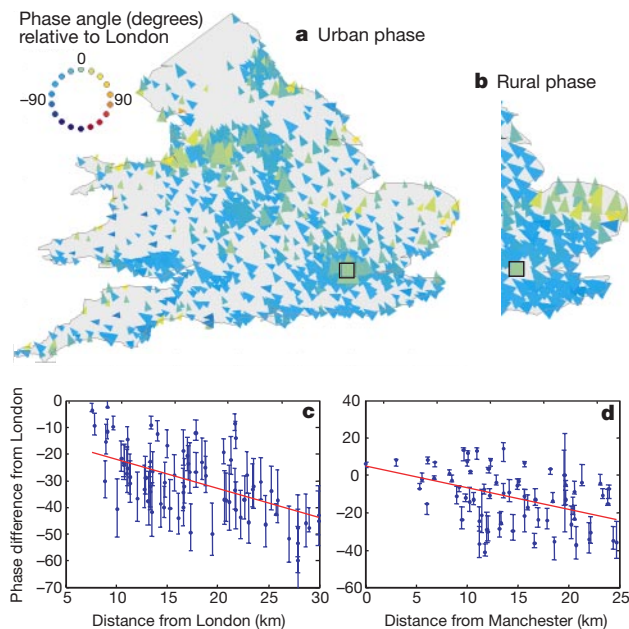


Figure 3 Phase differences from London for the full urban data set (954 locations). **a**, Urban mean phase differences from London for the pre-vaccination (1950–66) data, based on wavelet spectra (see Methods). Phases are coded by colour and angle as shown on the circular legend (see the Supplementary Information for more details). London is represented by a square. **b**, Phase differences from London of rural districts between London and Norwich; colour coding is as in **a**. **c**, Mean urban phase difference from London for 1950–66 in the London region, as a function of distance from the capital city. Within 30 km of London, there is a significant correlation of phase difference and distance ($r = -0.59$, 99% bootstrap limits: -0.75 to -0.39); the error bars are 99% bootstrapped confidence limits. **d**, As **c**, but showing phase difference from London in the Manchester area.

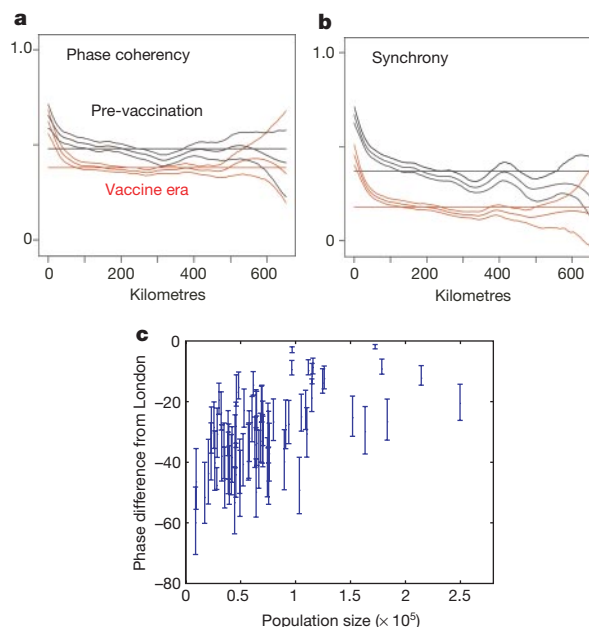


Figure 4 Pre-vaccination (black) and vaccine era (red). **a**, Spatial phase coherency function (a) and spatial synchrony of measles cases (b) with 95% confidence limits (see Methods). The horizontal bars represent the grand average phase coherence and grand average synchrony split into the pre-vaccination and vaccination eras. **c**, Phase difference from London as a function of population size for towns within 30 km of the capital (compare with Fig. 3c). Partial correlation analysis shows that phase difference is significantly and independently related both to distance from London ($r = -0.5$, $P < 0.001$; Fig. 3c) and population size ($r = 0.46$, $P < 0.0001$).

We investigated several different spatially coupled models for measles dynamics to understand the nature and cause of the hierarchical epidemic waves revealed by the wavelet analysis. Our most successful modelling framework to date builds on Bartlett's chain models¹⁶, as refined in the time series SIR (TSIR) model^{21,22,49}. The empirical waves are characterized by certain large cities (London and Manchester/Leeds) leading their local environs by a few to several weeks (Fig. 3c), with more complex relationships in more distant local areas. On the basis of the TSIR framework, we set up a simple coupled-map lattice, where at time $n + 1$ the epidemic intensity, λ_{n+1} , is given as:

$$\lambda_{n+1} = \beta_n S_n (I_n + \theta_n)^\alpha \quad (1)$$

Here, β_n is a seasonal transmission rate, α is the mixing rate²¹, S_n and I_n are respectively the number of local susceptible and infectious individuals in the previous characteristic time step (equal to 2 weeks, the approximate generation time for measles infection); θ_n is the (stochastic) contact rate with infectives, $\bar{I}_n$, in the neighbouring populations in the previous characteristic time step. We assume this rate to be Poisson distributed, θ_n is distributed as $Po(m\bar{I}_n)$, where m is the spatial coupling rate. According to the TSIR model, the epidemic birth-and-death process is realized according to a negative-binomial distribution⁴⁹ so that

$$I_{n+1} \text{ is distributed as } \text{NegBin}(\lambda_{n+1}, I_n) \quad (2)$$

The balance equation for the susceptibilities is $S_{n+1} = S_n + B - I_{n+1}$, where B is the birth rate. The parameter values closely match the best estimates for London and the estimated laws for how transmission scales with community size⁴⁹ (see Methods).

We envisage a human population distributed between a central city, composed of more tightly linked boroughs, and peripheral towns. Fig. 5a and b is generated from a linear map of a city with nine boroughs (shaded; for which m is arbitrarily set at 0.2) surrounded by the ten peripheral towns (for which m is set at 0.1):



In the above map, we assume that the population size is the same at all locations—this is not critical, but it makes interpretation easier. We contrast three situations.

In the first (Fig. 5a), each location is assumed to be below the CCS with 200,000 inhabitants. In an isolated town of this size, we would expect Type II dynamics²⁷—regular biennial outbreaks, but with occasional local extinctions in the troughs. The dynamics from our spatial model show:

- (1) High correlations between the tightly coupled boroughs (marked by black circles); the mean correlation is 0.79.
- (2) Epidemics in the periphery lag behind those of the centre; the lag depends on the distance from the core (so that the mean correlations

focus on the relatively simple spatial hierarchy surrounding London in the pre-vaccination era (Figs 3c and 4c), in order to understand the essence of the epidemic waves. In this region, the wave moves fairly uniformly away from London (Fig. 3c). However, there is evidence that, in addition to distance, the local population size is important—smaller centres lag more than larger ones (Fig. 4c). Further analysis shows that both distance from London and population size contribute significantly to the epidemic lag (Fig. 4 legend).

This set of observations prompts the following conceptual model for the generation of travelling waves in childhood epidemics. Consider two epidemiologically coupled towns, with biennial measles cycles that are roughly in phase (for example, see Fig. 5a). Assume further that one town is large—above the critical community size (CCS) for measles persistence of around 300,000 (ref. 35)—

of the first through fifth periphery are 0.72, 0.63, 0.54, 0.50 and 0.45 respectively).

(3) The drop in correlation across the periphery is associated with a progressive time delay in peak incidence (by 1, 1.5, 2, 2.5 and 4 biweeks as we move from the nearest to the furthest peripheral town). This phase lag corresponds crudely to that seen in the data.

Figure 5a–c shows the mean of 10 biennia. Raw epidemic time series from the model give qualitatively similar results, as do the associated wavelet phases of the simulations (see the Supplementary Information). We stress that, whereas spatial coupling rates have been chosen somewhat arbitrarily and for illustrative purposes, all other parameters are estimated from data⁴⁹.

In the second case (Fig. 5b), each unit is above the CCS, with 3 million inhabitants. Consequently, each location exhibits persistent (Type I) biennial dynamics²⁷. With local persistence at all locations, all are locked onto much more synchronized attractors than in Fig. 5a. The mean correlation between the boroughs is 0.98, and that between the boroughs and the periphery is 0.97. (A strictly deterministic model also generates similar results.) This relative phase locking contrasts markedly with the wave in Fig. 5a. See the Supplementary Information for a wavelet phase analysis of the model output.

Wave-like spatio-temporal behaviour might also result if there were a hierarchical trend of reduced infection rate as we move from large cities to smaller centres^{15,50}. However, in practice, the basic reproductive ratio of measles infection, R_0 , was relatively constant across the towns and cities of England and Wales⁴⁹, contradicting this alternative explanation. In fact, our simple coupling assumptions caused a slight increase in R_0 in the central city compared to the periphery, but its effect is small compared to the hierarchical invasion waves. The Supplementary Information explores these issues in more detail.

Finally, when we increase the arena of Fig. 5a to encompass a greater periphery (more small towns), we occasionally obtain travelling waves across the whole arena, as before. However, frequently we see a more complicated picture in which the more distant locations lock onto the other coexisting attractor (if the core peaked in even years, the periphery would peak in odd years, and *vice versa*). This is illustrated in Fig. 5c, where we use nine peripheral towns on either side of the city, rather than five. The simulation illustrates a very similar wave to Fig. 5a in the five inner towns of the periphery. By contrast, the outer four towns fall onto the opposite biennial attractor in this simulation. Such a reversal of phase can be stable for several epidemics and may give a qualitative explanation of the behaviour of Norwich and its environs (Fig. 1); isolation—and being near an edge (here the coast; see Supplementary Information)—increase the tendency for such jumps. Of course, towns are not connected only locally in terms of the movement of infection: an important area for future work is to consider how long-range 'jumps' of infection⁶ obscure local waves and move the system closer to mean-field behaviour. We shall also consider how regional spatial structure interacts with population size and birth rate to influence the hierarchical pattern of waves.

and that the other town is below the CCS:

(1) After a large epidemic, susceptible densities build up to the deterministic threshold, above which another epidemic can happen¹⁸.

(2) In the large town, measles is endemic throughout the inter-epidemic trough, so that a new epidemic occurs as soon as the effective reproductive ratio of infection exceeds unity¹⁸; this threshold is determined by the accumulation of susceptible children, modified by seasonally varying transmission rates associated with the school year.

(3) By contrast, in the small town, infection goes extinct locally after an epidemic; therefore, another epidemic cannot happen until an infective 'spark' is received, generally originating in a larger (endemic) community.

This reasoning prompts the following hypotheses:

(1) The spatial mosaic of large and small places is responsible for the lag in epidemics—we propose that having a group of small towns surrounding a large conurbation generates the hierarchical epidemic waves.

(2) As a corollary, coupling large centres (all above the CCS) should generate highly synchronized epidemics (as a result of nonlinear phase locking of seasonally forced oscillators)³¹.

(3) Finally, weakly coupled or distant centres should have a stronger tendency than nearer towns to move onto the 'opposite' biennial attractor from the main conurbation.

We tested these conjectures using a mechanistic model for the spread of measles across a linear array of locally coupled sites (Box 2). The model generates a picture in tight agreement with our conceptual scenario. First, if we couple a central large city to an array of outlying towns below the CCS, we observe a spatio-temporal wave of infection. Epidemics in the outlying towns lag progressively behind those in the city (Fig. 5a). Note that dividing the city into tightly coupled 'boroughs' does not generate a lag within the large centre. Second, if we repeat the analysis by coupling a series of large communities all above the CCS, the result is near-perfect coherence, with no evidence of epidemic waves (Fig. 5b). Because epidemics do not suffer local extinction, and because all the cities experience the same seasonal forcing, no lags are generated. Thus, the observed travelling waves are best seen as repeated and very fast invasion waves, extending from endemic core areas into epidemic satellite regions.

Third, we examine the effects of relative epidemiological isolation by extending the array of distant, more loosely coupled communities (Fig. 5c). As before, synchronized major epidemics with superimposed spatio-temporal waves appear for small communities close to the large one. However, epidemics in more distant towns can drop onto the opposite biennial attractor. This is preliminary evidence that relative isolation may be a source of the unusual 'even year' behaviour of pre-vaccination measles epidemics in Norwich and its environs (Fig. 2), during parts of the pre-vaccination era.

Our study raises a number of issues. We illustrate how wavelets can be used for analysing both non-stationary time series and spatio-temporal patterns in ecology and population biology. Spatial and temporal non-stationarity is the norm in ecology. It may be caused by a number of factors, including anthropogenic influences and biological evolution, as well as more esoteric dynamical effects, such as coexisting attractors^{20,36} and intermittent periodicity in chaos³⁷. Wavelets are emerging as essential tools in the study of intermittent processes in the physical sciences and experimental biology²⁶. Here we have illustrated a basic application of wavelet analysis in population dynamics, complementing previous applications to spatial transect patterns³⁸; there should be significant scope for further careful ecological application of wavelets. The wavelet phase analysis also provides an unusual method for analysing ecological travelling waves. Although other statistical models have been successfully developed^{11,39}, wavelet phase angles are naturally suited to allowing for variations in cycle period and amplitude.

In addition, the non-stationarity that we demonstrate in the measles series is due mainly to variations in the recruitment of susceptible individuals, driven by birth rates and especially by the onset of vaccination^{22,32}. Mass immunization caused a dramatic fall in susceptible recruitment rates and a marked reduction in the spatial correlation and coherence of epidemics: this may have important consequences for the design of vaccination programmes^{19,32}. The implications of hierarchical ('core-satellite') epidemic dynamics for the control of established and emergent diseases are important areas for future research.

Finally, our analysis demonstrates prominent repeated spatio-temporal waves, superimposed on the well-known epidemic time series of measles in England and Wales. This echoes, on a more extensive spatio-temporal scale, seminal work on the hierarchical spatial diffusion of measles epidemics¹⁴. The simplicity of the

measles dynamical 'clockwork' and the intricacy of the human demographic record allow us this unusual opportunity to quantify the impact of spatio-temporal heterogeneities on epidemic dynamics. The clearest temporal transition is the effect of mass vaccination. On a shorter timescale, seasonal variation in infection rate is a major driver of the dynamics^{40,41}. Classical reaction-diffusion models of host-natural enemy systems often predict recurring travelling waves in homogeneous (non-hierarchical) systems¹. However, such waves are precluded for measles, because of the strongly synchronizing effect of seasonally forced transmission³¹. Previous theory^{31,33,42} indicates that seasonally driven epidemics will either be completely synchronized across large coupled centres or more irregular in small centres buffeted by demographic noise.

This prediction is at odds with the observed waves (Figs 3 and 4),

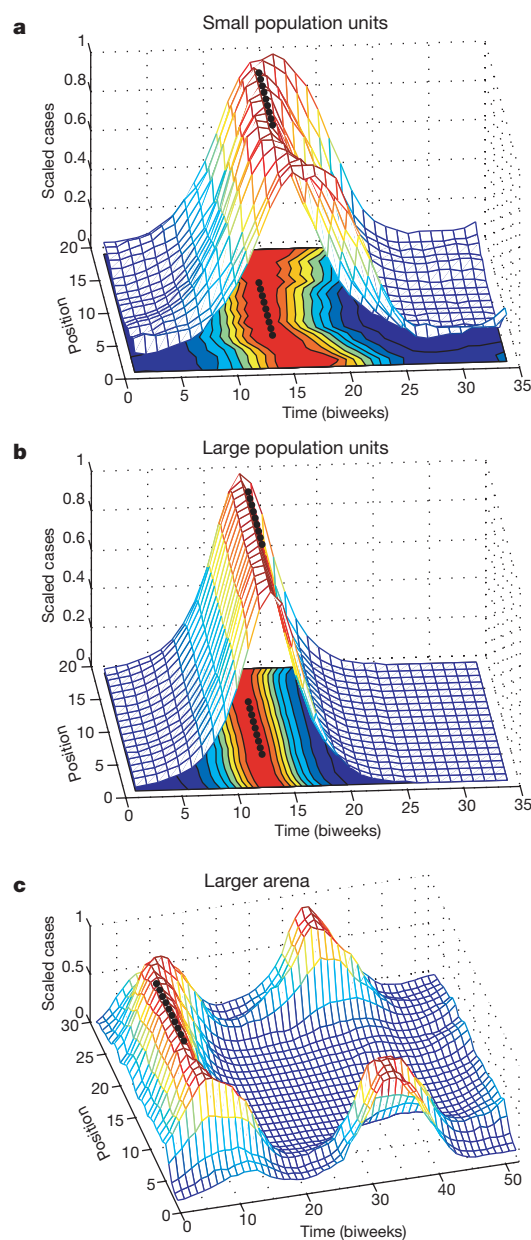


Figure 5 Average biennial behaviour of a spatially coupled measles model (see Box 2 for details). **a**, Coupling relatively small peripheral population units (below the critical community size, CCS, for measles persistence) to a central 'city'. **b**, Coupling larger populations (each above the CCS) to the city. **c**, As in **a**, but extending the size of the linear array of populations.

leading us to propose a different underlying mechanism—measles waves arise as repeated invasions from endemic core areas to the periphery. The coupling of large and small centres may thus be the critical feature that generates waves in these seasonally forced systems. This mechanism has analogies with the operation of the ‘rescue effect’ in core–satellite metapopulations⁴³ and also with forest fire dynamics⁴⁴. Forest-fire-like epidemic dynamics have previously been explored in the irregular, epidemic measles outbreaks observed on isolated small islands⁴⁵. This work provides evidence of such phenomena in an endemic epidemiological context. □

Methods

The data

We use weekly official measles notifications and associated demographic data for England and Wales^{15,19,46}. For the pre-vaccination era, before 1966, weekly measles case data are available for 945 cities and towns and 457 rural districts; these data sets are the basis of the pre-vaccination wavelet analysis (Figs 2, 3a, b and 4c). In 1974, boundary changes agglomerated the spatial data into 354 administrative areas. To achieve consistent time series across both pre-vaccination and vaccine era, we have therefore binned the pre-vaccination data into the post-1974 boundaries (Figs 1 and 4a, b).

Wavelet time series analysis

See Box 1. All series are logarithm transformed (after adding a constant of one) to make them more sinusoidal, and subsequently scaled to zero mean and unit variance. Assume that we have a time series, I_n , $n = 0, \dots, T - 1$ (for consistency, n is used as the time index throughout the paper). We analyse temporal changes in the distribution of power at different scales s (approximately periods) using a Morlet wavelet function, $\psi_0(\eta) = \pi^{-1/4} \exp(i\omega_0\eta) \exp(-\eta^2/2)$, where ω_0 is the non-dimensional frequency here taken to be 6 (ref. 23) and $\eta = n/s$. The Morlet wavelet, ψ_0 , is a damped complex exponential (Box 1).

The continuous wavelet transform (CWT), $W_n(s)$, of the time series I_n ($n = 0, 1, 2, \dots$) is calculated as the convolution of I_n with a scaled and translated version of ψ_0 (ref. 23). For the Morlet wavelet, scale is approximately equal to the Fourier period, so that the lowest scale, s_0 , roughly corresponds to the maximum (Nyquist) frequency of 0.5 cycles per time step. The local wavelet power spectrum (Box 1), at time point n and scale s is then given by $|W_n(s)|^2$.

To minimize biases due to edge effects, the data were padded with zeros, up to the next-highest power of two²³. The ‘cone of influence’ (the parabola in Fig. 1b) is a reflection of a consequent loss in statistical power near the start and end of the series; the area below the parabola in Fig. 1b should be interpreted cautiously. The width of the cone also gives a rough lower limit for how wide a feature needs to be at a given scale for it to represent genuine cyclical behaviour, rather than a spike. We conduct significance tests using methods described and discussed in ref. 23.

To isolate the major biennial epidemics for the phase analysis, we use wavelet reconstruction²³ to partition the original series, in terms of the contribution of variation at different frequencies.

Phase relationships between time series

A wavelet transform based on a complex wavelet, such as the Morlet wavelet, has a phase angle, defined by

$$\Theta_n(s) = \arctan[\text{Im}[W_n(s)]/\text{Re}[W_n(s)]]$$

where $\text{Re}[W_n(s)]$ is the real, and $\text{Im}[W_n(s)]$ the imaginary part of $W_n(s)$. In practice, we use this formula to calculate the time series of phase angles (θ_n) associated with the wavelet transform of the reconstructed major epidemic cycle for each time series. These can be used as absolute phases (Fig. 2d) or phase differences (Fig. 2e); phases and phase differences are restricted to the range $\pm 180^\circ$ (see Fig. 2, legend).

Algorithms

The algorithms and notation used here are based on a practical guide to meteorological wavelet analysis²³. See also <http://paos.colorado.edu/research/wavelets/> for more background information. The website also includes a set of wavelet algorithms in Matlab, on which the analyses shown here are based.

Phase coherency and correlation functions

We calculate the phase coherency, ϑ_{ij} , between two cities, i and j , by considering the correlation between the time series of phase angles, $\theta_{i,n}$ and $\theta_{j,n}$:

$$\vartheta_{ij} = \sum_n (\theta_{i,n} - \bar{\theta}_i)(\theta_{j,n} - \bar{\theta}_j)/\sigma_i\sigma_j$$

where $\bar{\theta}$ represents the respective mean phases, and σ the respective standard deviation of the phases. We further define the phase coherency function, $\eta(d)$, which governs how the phase coherency relates to the distance, d , separating any two cities. Assuming that the phase coherency forms a stationary random field, then the phase coherency function (PCF) represents the spatial correlation function of that field. We estimate the PCF,

without making *a priori* assumptions about the functional form, by using the nonparametric covariance function (NCF)^{7,47,48}. The NCF uses either a kernel function or (as in our case) a cubic B-spline to estimate the underlying correlation function. We use a smoothing spline with 25 equivalent degrees of freedom (e.d.f.), and quantify the confidence envelopes for the functions using 500 bootstrap iterations (see ref. 48 for details of estimation and testing).

We measure epidemic synchrony between two populations as the Pearson correlation coefficient between the two variance-stabilized (square-root transformed) time series of incidence. We calculate the spatial correlation functions governing how this synchrony depends on distance—again, using the nonparametric spline covariance function.

Separate phase coherency and spatial correlation functions are calculated for the pre-vaccination (1954–64) and vaccination (1970–80) data, based on the 354 time series that span both eras.

The spatial TSIR model

We used the following parameters for the model in Box 2: $\alpha = 0.97$, and $\{\beta_n\}$ is an annually periodic function with the following values for the 26 biweeks of the year: 1.11×10^{-5} , 1.11×10^{-5} , 1.10×10^{-5} , 1.09×10^{-5} , 1.06×10^{-5} , 1.03×10^{-5} , 1.01×10^{-5} , 9.84×10^{-6} , 9.63×10^{-6} , 9.40×10^{-6} , 9.10×10^{-6} , 8.71×10^{-6} , 8.29×10^{-6} , 7.89×10^{-6} , 7.60×10^{-6} , 7.49×10^{-6} , 7.60×10^{-6} , 7.93×10^{-6} , 8.42×10^{-6} , 8.95×10^{-6} , 9.43×10^{-6} , 9.80×10^{-6} , 1.01×10^{-5} , 1.03×10^{-5} , 1.05×10^{-5} , 1.08×10^{-5} . These values are a smoothed version of the estimated parameters for the London time series⁴⁹. The biweekly birth rate, B , is set at 100 for small population units (Fig. 5a) and 2,000 for large ones (Fig. 5b).

Received 4 May; accepted 12 October 2001.

- Murray, J. D. *Mathematical Biology* (Springer, London, 1989).
- Neubert, M. G., Kot, M. & Lewis, M. A. Dispersal and pattern formation in a discrete-time predator–prey model. *Theor. Pop. Biol.* **48**, 7–43 (1995).
- Sherratt, J. A. Periodic travelling waves in cyclic predator–prey systems. *Ecol. Lett.* **4**, 30–37 (2001).
- Hassell, M. P., Comins, H. N. & May, R. M. Spatial structure and chaos in insect population dynamics. *Nature* **353**, 255–258 (1991).
- Sole, R. V., Valls, J. & Bascompte, J. Spiral waves, chaos and multiple attractors in lattice models of interacting populations. *Phys. Lett. A* **166**, 123–128 (1992).
- Jeltsch, F., Müller, M. S., Grimm, V., Wissel, C. & Brandl, R. Pattern formation triggered by rare events: Lessons from the spread of rabies. *Proc. R. Soc. Lond. B* **264**, 495–503 (1997).
- Bjørnstad, O. N., Ims, R. A. & Lambin, X. Spatial population dynamics: analysing patterns and processes of population synchrony. *Trends Ecol. Evol.* **14**, 427–432 (1999).
- Blasius, B., Huppert, A. & Stone, L. Complex dynamics and phase synchronization in spatially extended ecological systems. *Nature* **399**, 354–359 (1999).
- Mollison, D. Modeling biological invasions—chance, explanation, prediction. *Phil. Trans. R. Soc. Lond. B* **314**, 675–693 (1986).
- Bulmer, M. G. A statistical analysis of the 10-year cycle in Canada. *J. Anim. Ecol.* **43**, 701–718 (1974).
- Lambin, X., Elston, D. A., Petty, S. J. & MacKinnon, J. L. Spatial asynchrony and periodic travelling waves in cyclic populations of field voles. *Proc. R. Soc. Lond. B* **265**, 1491–1496 (1998).
- Ranta, E. & Kaitala, V. Travelling waves in vole population dynamics. *Nature* **390**, 456–456 (1997).
- Stenseth, N. C. *et al.* Common dynamic structure of Canada lynx populations within three climatic regions. *Science* **285**, 1071–1073 (1999).
- Cliff, A. D., Haggett, P. & Smallman-Raynor, M. *Measles: An Historical Geography of a Major Human Viral Disease from Global Expansion to Local Retreat, 1840–1990* (Blackwell, Oxford, 1993).
- Grenfell, B. T. & Bolker, B. M. Cities and villages: infection hierarchies in a measles metapopulation. *Ecol. Lett.* **1**, 63–70 (1998).
- Bartlett, M. S. in *Proc. Third Berkeley Symp. on Mathematical Statistics and Probability* Vol. 4, 81–109 (Univ. California Press, Berkeley, 1956).
- Schenzle, D. An age-structured model of pre- and post-vaccination measles transmission. *IMA J. Math. Appl. Med. Biol.* **1**, 169–191 (1984).
- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
- Grenfell, B. T. & Harwood, J. (Meta)population dynamics of infectious diseases. *Trends Ecol. Evol.* **12**, 395–399 (1997).
- Earn, D. J. D., Rohani, P., Bolker, B. M. & Grenfell, B. T. A simple model for complex dynamical transitions in epidemics. *Science* **287**, 667–670 (2000).
- Finkenstädt, B. F. & Grenfell, B. T. Time series modelling of childhood diseases: a dynamical systems approach. *J. R. Stat. Soc. C* **49**, 187–205 (2000).
- Grenfell, B. T., Bjørnstad, O. N. & Finkenstädt, B. F. Dynamics of measles epidemics. II. Scaling noise, determinism and predictability with the time series SIR model. *Ecol. Monogr.* (in the press).
- Torrence, C. & Compo, G. P. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**, 61–78 (1998).
- Nason, G. P. & von Sachs, R. Wavelets in time series analysis. *Phil. Trans. R. Soc. Lond. A* **357**, 2511–2526 (1999).
- Daubechies, I. *Ten Lectures on Wavelets* (Society for Industrial and Applied Mathematics, Philadelphia, 1992).
- Percival, D. B. & Walden, A. T. *Wavelet Methods for Time Series Analysis* (Cambridge Univ. Press, Cambridge, 2000).
- Bartlett, M. S. Measles periodicity and community size. *J. R. Stat. Soc. A* **120**, 48–70 (1957).
- Anderson, R. M., Grenfell, B. T. & May, R. M. Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. *J. Hyg.* **93**, 587–608 (1984).
- Chatfield, C. *The Analysis of Time Series: An Introduction* (Chapman and Hall, London, 1996).
- Finkenstädt, B. F. & Grenfell, B. T. Empirical determinants of measles metapopulation dynamics in England and Wales. *Proc. R. Soc. Lond. B* **265**, 211–220 (1998).
- Lloyd, A. L. & May, R. M. Spatial heterogeneity in epidemic models. *J. Theor. Biol.* **179**, 1–11 (1996).
- Earn, D., Rohani, P. & Grenfell, B. T. Spatial dynamics and persistence in ecology and epidemiology. *Proc. R. Soc. Lond. B* **265**, 7–10 (1998).

33. Bolker, B. M. & Grenfell, B. T. Impact of vaccination on the spatial correlation and dynamics of measles epidemics. *Proc. Natl Acad. Sci. USA* **93**, 12648–12653 (1996).
34. Cliff, A. D., Haggett, P., Stroup, D. F. & Cheney, E. The changing geographical coherence of measles morbidity in the United States, 1962–88. *Stat. Med.* **11**, 1409–1424 (1992).
35. Bartless, M. S. The critical community size for measles in the U.S. *J. R. Stat. Soc. A* **123**, 37–44 (1960).
36. Grenfell, B. T., Bolker, B. M. & Kleczkowski, A. Seasonality and extinction in chaotic metapopulations. *Proc. R. Soc. Lond. B* **259**, 97–103 (1995).
37. Kendall, B. E., Schaffer, W. M. & Tidd, C. W. Transient periodicity in chaos. *Phys. Lett. A* **177**, 13–20 (1993).
38. Bradshaw, G. A. & Spies, T. A. Characterizing canopy gap structure in forests using wavelet analysis. *J. Ecol.* **80**, 205–215 (1992).
39. Bjørnstad, O. N. & Bascompte, J. Synchrony and second order spatial correlation in host-parasitoid systems. *J. Anim. Ecol.* **70**, 924–933 (2001).
40. Schaffer, W. M. & Kot, M. Nearly one dimensional dynamics in an epidemic. *J. Theor. Biol.* **112**, 403–427 (1985).
41. Schwartz, I. B. in *Biomedical Modelling and Simulation* (eds Eisenfeld, J. & Levine, D. S.) 201–204 (Scientific, New York, 1989).
42. Bolker, B. M. & Grenfell, B. T. Space, persistence and the dynamics of measles epidemics. *Phil. Trans. R. Soc. Lond. B* **348**, 309–320 (1995).
43. Hanski, I. & Gyllenberg, M. Two general metapopulation models and the core-satellite hypothesis. *Am. Nat.* **142**, 17–41 (1993).
44. Bak, P. *How Nature Works—The Science of Organised Criticality* (Oxford Univ. Press, Oxford, 1977).
45. Rhodes, C. J. & Anderson, R. M. Power laws governing epidemics in isolated populations. *Nature* **381**, 600–602 (1996).
46. Office of Population Censuses and Surveys Registrar General's Weekly Return for England and Wales (Her Majesty's Stationery Office (HMSO), London, 1944–94).
47. Hall, P. & Patil, P. Properties of nonparametric estimators of autocovariance for stationary random fields. *Prob. Theor. Relat. Fields* **99**, 399–424 (1994).
48. Bjørnstad, O. N. & Falck, W. Nonparametric spatial covariance functions: estimation and testing. *Environ. Ecol. Stat.* **8**, 53–70 (2001).
49. Bjørnstad, O. N., Finkenstädt, B. & Grenfell, B. T. Dynamics of measles epidemics. I. Estimating scaling of transmission rates using a time series SIR model. *Ecol. Monogr.* (in the press).
50. May, R. M. & Anderson, R. M. Spatial heterogeneity and the design of immunization programs. *Math. Biosci.* **72**, 83–111 (1984).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com>).

Acknowledgements

We thank S. Cornell, T. Coulson, J. Gog, M. Keeling, A. Lloyd, G. Nason, P. Rohani, C. Torrence and C. Williams for helpful discussions. B.T.G. and J.K. were supported by the Wellcome Trust. O.N.B. was supported by the National Center for Ecological Analysis and Synthesis (a Centre funded by an NSF grant, the University of California Santa Barbara, and the State of California) and the Norwegian Science Foundation.

Correspondence and requests for materials should be addressed to B.T.G. (e-mail: b.t.grenfell@zoo.cam.ac.uk).

Temporal evolution of the electric field accelerating electrons away from the auroral ionosphere

G. T. Marklund*, N. Ivchenko*, T. Karlsson*, A. Fazakerley†, M. Dunlop‡, P.-A. Lindqvist*, S. Buchert§, C. Owen†, M. Taylor†, A. Vaivalds§, P. Carter†, M. André§ & A. Balogh‡

* Division of Plasma Physics, Alfvén Laboratory, KTH, Royal Institute of Technology, SE 10044 Stockholm, Sweden

† University College, London, Mullard Space Science Laboratory, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK

‡ Imperial College, Blackett Laboratory, Space and Atmospheric Physics Group, London SW7 2BW, UK

§ Swedish Institute of Space Physics, Ångströmlaboratoriet, Box 534, SE 751 21 Uppsala, Sweden

The bright night-time aurorae that are visible to the unaided eye are caused by electrons accelerated towards Earth by an upward-pointing electric field^{1–3}. On adjacent geomagnetic field lines the reverse process occurs: a downward-pointing electric field accelerates electrons away from Earth^{4–11}. Such magnetic-field-aligned electric fields in the collisionless plasma above the auroral ionosphere have been predicted¹², but how they could be maintained is still a matter for debate¹³. The spatial and temporal behaviour of the electric fields—a knowledge of which is crucial to an understanding of their nature—cannot be resolved uniquely by single satellite measurements. Here we report on the first observations

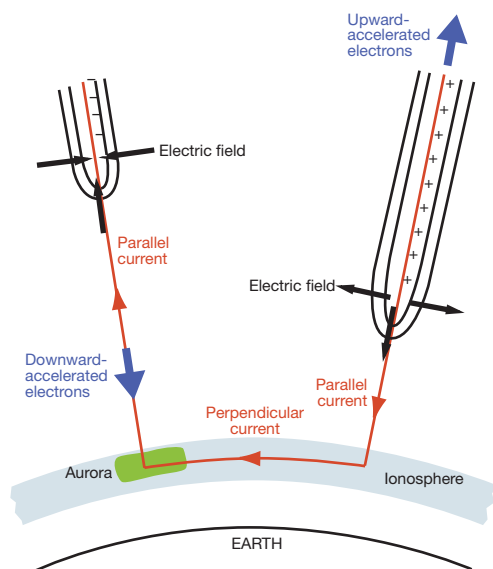


Figure 1 Acceleration structures (black) within the auroral current circuit (red). This diagram shows a negatively charged potential structure (indicated by equipotential contours) representative of the aurora (left) and a positively charged potential structure representative of the auroral return current (right). The two branches of the equipotential contours close at typical altitudes of 5,000–8,000 km in the upward-current region above the aurora and at 1,500–3,000 km in the auroral return-current region, respectively, above which they extend to very high altitudes along the geomagnetic field, forming the characteristic U-shape. The field-aligned currents are carried by the downward and upward accelerated electrons, respectively (blue). Together with the ionospheric closure current and the magnetospheric generator these form the complete auroral current circuit. The figure shows a north–south section through the structures that usually extend several hundreds of kilometres in the east–west direction.

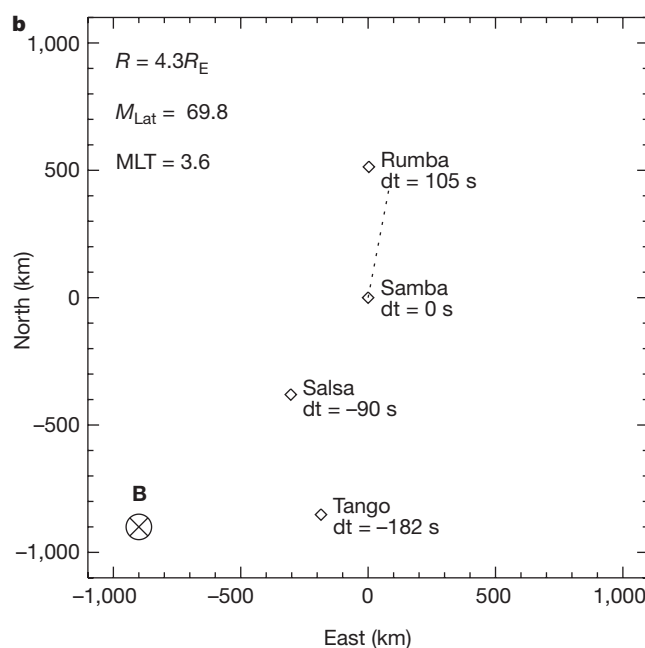
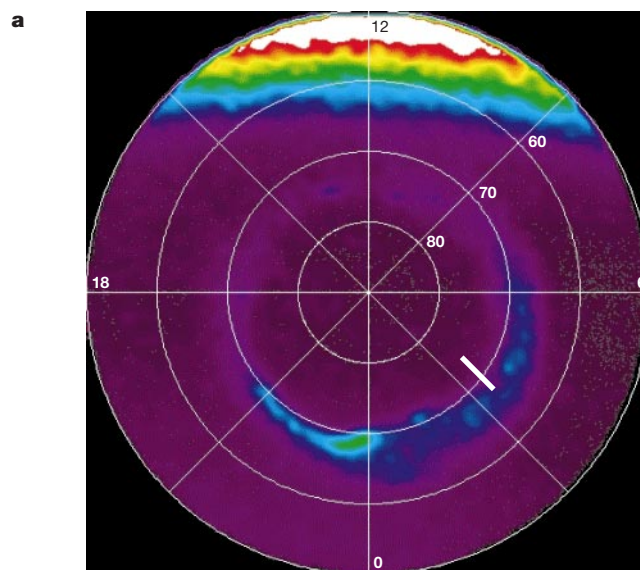


Figure 2 Location and configuration of the Cluster satellite formation passing over the Northern Hemisphere auroral oval. It passed between 4:20 and 4:40 UT on 14 January 2001, in the recovery phase of a weak substorm, peaking at 04:00 UT. **a**, The nightside auroral oval as observed by the ultraviolet imager on the IMAGE satellite at 04:30 UT. The post-midnight auroral oval forms a relatively homogeneous diffuse band between 65 and 70 degrees magnetic latitude. The ionospheric projection of the Cluster orbit (between 4:20 and 04:45 UT) is shown by the white line intersecting the auroral oval near 3:00 magnetic local time. **b**, Cluster satellite configuration on 14 January 2001. Relative positions and time differences (dt) between the Cluster 1, 2, and 4 spacecraft with respect to the reference Cluster spacecraft 3 (Samba) in a plane transverse to the magnetic field. The four Cluster spacecraft are aligned nearly as pearls on a string, with the separation transverse to the velocity vector being small, and follow in the sequence 1–3–2–4. The direction of motion of Samba projected onto this plane is shown by the dotted line. **B** denotes the direction of the geomagnetic field, R is the geocentric distance of the Samba spacecraft expressed in units of the Earth radius (R_E), M_{Lat} denotes magnetic latitude, and MTL denotes magnetic local time.

by a formation of identically instrumented satellites crossing a beam of upward-accelerated electrons. The structure of the electric potential accelerating the beam grew in magnitude and width for about 200 s, accompanied by a widening of the downward-current sheet, with the total current remaining constant.

The 200-s timescale suggests that the evacuation of the electrons from the ionosphere contributes to the formation of the downward-pointing magnetic-field-aligned electric fields. This evolution implies a growing load in the downward leg of the current circuit, which may affect the visible discrete aurorae.

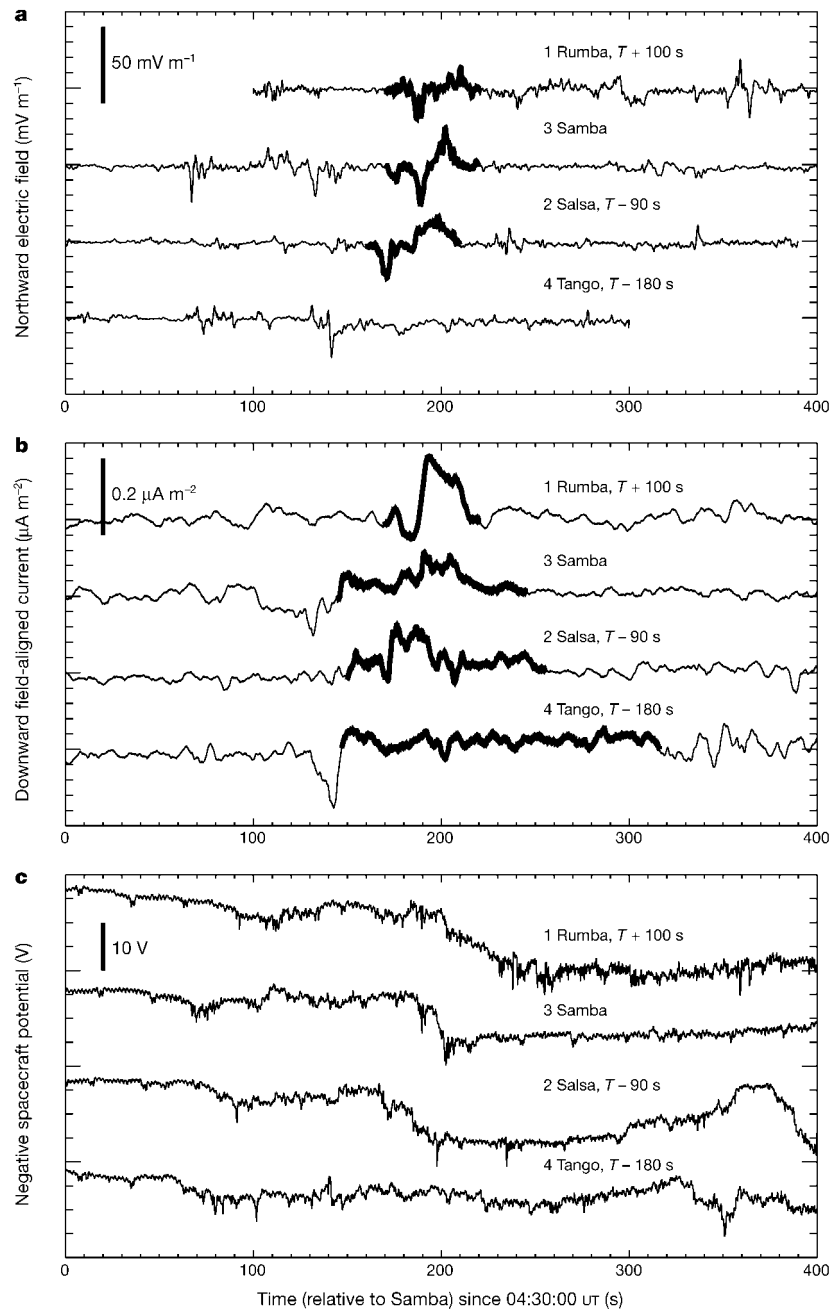


Figure 3 Evolution of the electric field, magnetic field-aligned current, and plasma density between the four crossings. The time axes for all satellites except for Samba (T) are shifted to account for the time delay between the satellites. **a**, Geomagnetically northward electric field component normal to the magnetic field, derived from measurements by the Electric Field and Wave (EFW) instrument onboard the four satellites. The bipolar structure (a southward excursion followed by a northward excursion) is highlighted in bold in the centre of each panel. It is seen to grow both in latitudinal width (from 15 km to 25 km projected to the ionospheric level) and magnitude between the first three crossings and then fade away by the time of the last crossing. Electric potentials obtained by integration along the orbit increase from 0.4 kV on Rumba, to 1.7 kV on Samba, to over 2 kV on Salsa. **b**, Field-aligned current derived from the geomagnetic east-west variation

of the magnetic field, measured by the Flux Gate Magnetometer (FGM) instrument. A downward-current sheet is co-located with the bipolar structures. The total current integrated over the sheet is roughly constant, but the current density decreases between the consecutive crossings, as the current sheet widens. The most significant widening occurs between the last two crossings. **c**, Negative of the spacecraft potential, as measured by the EFW instrument. The spacecraft potential depends primarily on the ambient plasma density, and is thus a monitor thereof. The electric-field structure is co-located with a density gradient in the first three crossings, separating a region of higher and less variable densities southward of the structure from a region of low and variable densities to the north. At the time of the fourth crossing the density gradient is no longer seen.

In situ observations of electric fields, field-aligned currents, and energetic electrons and ions made by numerous satellites above and below the regions where the auroral acceleration takes place have been used to confirm^{1–3} that the quasi-static electric potential structures formed above the aurora are typically U-shaped and negatively charged, as first suggested in ref. 14 on the basis of studies of the motion of auroral forms. These structures are aligned in the geomagnetic east–west direction, and are located in the region of upward-field-aligned current with an upward electric field (auroral acceleration region) at typical altitudes of 5,000–8,000 km (Fig. 1, left). A satellite moving poleward above the auroral acceleration region will measure the typical bipolar signature of a converging electric field and an upward beam of ions having a characteristic energy corresponding to the parallel potential drop below the satellite, whereas a satellite below this potential drop will observe accelerated auroral electrons and a much reduced electric field. The Swedish satellite Freja, with an apogee at 1,700 km, well below the lower limit of the auroral acceleration region, observed very intense (of the order of 1 V m^{-1}) and narrow (a few km) diverging electric fields^{7,8} accompanied by upward fluxes of energetic electrons (a few keV)⁹ on field lines adjacent to the aurora, where the downward current flows. These unexpected observations suggested that positively charged potential structures develop there (Fig. 1, right),

forming the counterpart to the negative potential structures of the aurora. Now there is conclusive evidence, particularly from observations by the satellite FAST at higher altitudes around 4,000 km, to confirm this picture¹⁰. A few theories of how to maintain such a structure with a downward electric field have been suggested^{4,6,15–17}. The Polar satellite, exploring the region at altitudes high above the acceleration region (19,000–38,000 km), has detected electric fields occasionally exceeding 100 mV m^{-1} (refs 18 and 19).

We need to understand the temporal behaviour of these structures to reveal by what mechanism they are formed. However, despite the excellent quality of the data from auroral missions such as Freja, FAST and Polar the spatio-temporal ambiguity inherent in these measurements prevented any conclusive answers. The European Space Agency mission Cluster^{20,21}, with four identically instrumented spacecraft (spacecraft 1, 2, 3 and 4 named Rumba, Salsa, Samba and Tango, respectively) in orbit since summer 2000, is the first ambitious and carefully planned attempt to carry out multipoint *in situ* measurements in the near-Earth space environment. Such measurements should be capable of solving this problem.

The Cluster observations presented here are from a poleward crossing of auroral field lines at an altitude of 22,000 km, where the

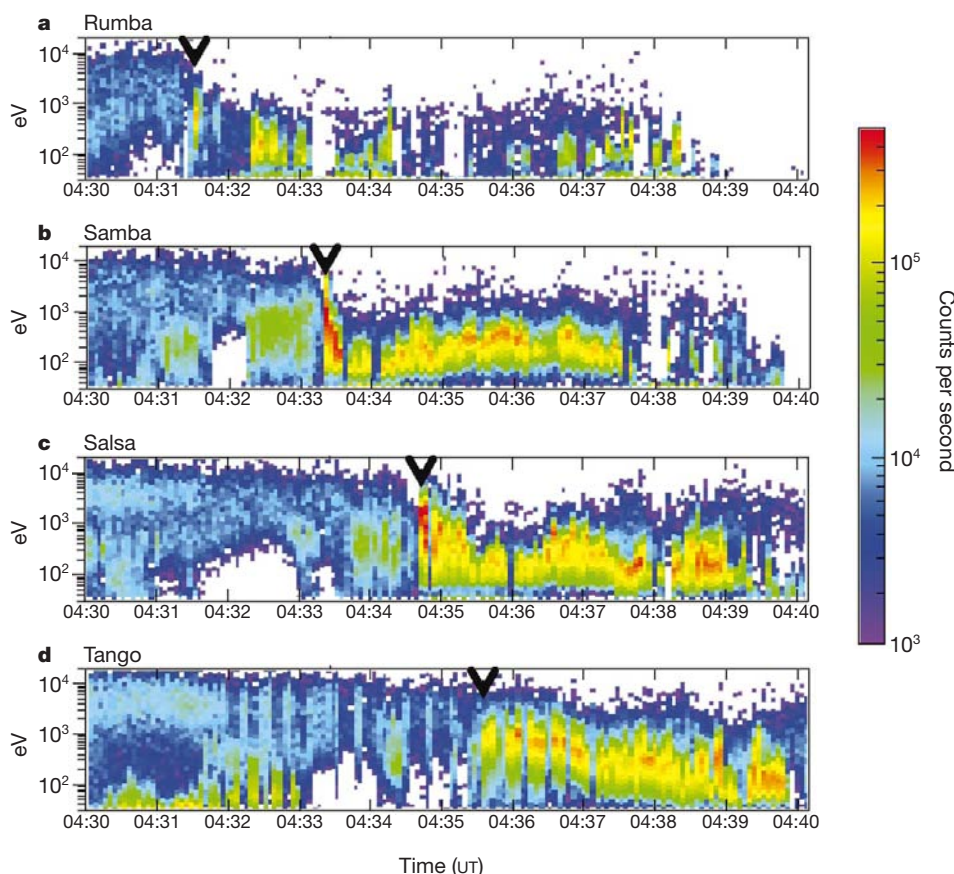


Figure 4 Upward-accelerated electron beam: evolution of energy and width. **a–d**, Time-energy spectrograms show electrons moving anti-parallel to the magnetic field as measured by the Plasma Electron and Current Experiment (PEACE) instrument between 4:30 and 4:40 UT, 14 January 2001 (the full pitch angle distribution is not shown here). The arrowheads at the top of the panels indicate the encounters with the bipolar-electric-field structures, occurring at a transition between two characteristically different electron populations, a high-energy isotropic electron population characteristic of the central plasma sheet to the south, and a lower-energy population characterized by intermittent, bi-directional, and field-aligned electron bursts, typical of the plasma sheet boundary

layer to the north, and collocated with an upward more energetic electron beam. The beam (**b**) first increases in energy from about 0.7 keV on Rumba (**a**), to 1.2 keV on Samba (**b**), to 2 keV on Salsa (**c**), but at the time of the Tango crossing (**d**) the energy has dropped to 0.8 keV and the beam becomes bi-directional. High-resolution data (not presented here) show that the width of the beam on spacecraft 1, 3, and 2 was roughly the same as the width of the potential structure. The false colour code represents the number flux of electrons in counts per second ranging from 10^3 (dark blue) to 5×10^5 (red).

four spacecraft moved along essentially the same orbit separated by 100 s (Fig. 2). During this auroral crossing spacecraft 1, 3 and 2 observed a diverging, bipolar electric field (Fig. 3) changing from a southward to a northward direction. Together with a dominant east–west magnetic-field variation this implies that the structure was sheet-like and extended in the east–west direction. It grew in latitudinal width and magnitude, corresponding to an increase in the positive electric potential peak of the structure from 0.4 kV on Rumba, to 1.7 kV on Samba, to over 2 kV on Salsa. Accompanying this structure was an upward uni-directional electron beam, increasing in energy (from about 0.7 keV on Rumba, to 1.2 keV on Samba, and to 2.0 keV on Salsa), as well as in width (Fig. 4). However, when Tango reached the position of the structure, the electric field structure and the unidirectional upward electron beam had vanished. The electric field structure and associated upward electron beam are located in the transition region between the central plasma sheet and the plasma sheet boundary layer. The region of diffuse aurora seen in the post-midnight auroral oval (Fig. 2) corresponds to the central plasma sheet electron precipitation, whereas the darker region just poleward of it corresponds to the plasma sheet boundary layer. This transition is also indicated by the gradient in the plasma density, which first steepens as the electric field increases, then becomes less steep as the electric field structure widens, and finally fades away together with the electric field at the time of the last crossing. The field-aligned current associated with the structure is downward, as shown in Fig. 3b. The total downward current integrated over the structure remains roughly constant during the crossings by the four spacecraft, whereas the width of the current sheet is seen to increase, in particular between the two last crossings.

The consistency between the evolution of the parallel acceleration potential (inferred from the energy of the upward electron beam) and that of the perpendicular potential (integrated from the electric field) indicates that the Cluster spacecraft were crossing the upper reaches of a U-shaped positive potential structure. The data show that this positive potential structure (diverging electric field structure) in the downward current region extends to altitudes greater than 20,000 km and that it grows in size and intensifies over a time period of a few hundred seconds, after which it fades away accompanied by a drop in the characteristic energy of the electron beam, a significant widening of the downward current sheet, and no distinct plasma density gradient. It has been shown by numerical simulations that this time period is comparable to the time it takes to evacuate the ionosphere electrons over the region of the downward current sheet²². This evacuation, accomplished by upward-moving electrons and by transport of ions perpendicular to the geomagnetic field in the lower ionosphere, implies an increasing load over the downward-field-aligned current branch of the auroral current circuit. From Freja observations⁹ we know that both the occurrence and the magnitude of the diverging electric fields peak for minimum ionospheric conductivity conditions, near winter solstice and local midnight⁹. Thus, it is possible that the evolution of the potential structure, with its region of a parallel electric field predominantly above the ionosphere, is closely related to the formation of such ionospheric plasma density holes, although the nature of this relationship remains to be understood. Temerin and Carlson¹⁶ have presented a model of the current–voltage relationship in the return current region based on quasi-neutrality. This model, however, does not include the plasma evacuation, which we believe is an essential feature of this region. A possible explanation for the fading of the potential structure by the time Tango reached its location may be found in the sudden widening of the return current sheet seen at this time (Fig. 3b). The current carriers now become available from a much wider region, so that the current can easily be maintained. This again agrees with the results from the numerical simulation of the return current dynamics that as the plasma density hole develops, charge carriers are collected from

regions further and further away from the core of the return current flux tube.

The negatively charged auroral potential structure and the positively charged return current structure are connected to each other in a common current circuit (Fig. 1), driven by an external generator^{23,24}. This generator may be connected with the diverging bipolar-electric-field structure observed, which could reflect a strong velocity shear, imposed by the magnetosphere at the interface between the central plasma sheet and the boundary plasma sheet. Whether the generator is connected along the magnetic field to the structure or not, the evolving potential structure and the associated ionospheric density hole form a growing load which influences the whole circuit, either through a direct feedback or through load redistribution, and thus it may have an impact on the ordinary visible aurora. □

Received 14 November; accepted 19 November 2001.

1. Reiff, P. H. *et al.* Determination of auroral electrostatic potentials using high- and low-altitude particle distributions. *J. Geophys. Res.* **93**, 7441–7465 (1988).
2. Block, L. P. & Fälthammar, C.-G. The role of magnetic field-aligned electric fields in auroral acceleration. *J. Geophys. Res.* **95**, 5877–5888 (1990).
3. McFadden, J. P., Carlson, C. W. & Ergun, R. E. Microstructure of the auroral acceleration region as observed by FAST. *J. Geophys. Res.* **104**, 14453–14480 (1999).
4. Chiu, Y. T., Newman, L. & Cornwall, J. M. On the structures and mapping of auroral electrostatic potentials. *J. Geophys. Res.* **86**, 10029–10037 (1981).
5. Klumpar, D. M. & Keikkilä, W. J. Electrons in the ionospheric source cone: Evidence for runaway electrons as carriers for downward Birkeland currents. *Geophys. Res. Lett.* **9**, 873–876 (1982).
6. Gorney, D. J., Chiu, Y. T. & Croley, D. R. Jr Trapping of ion conics by downward parallel electric fields. *J. Geophys. Res.* **90**, 4205–4210 (1985).
7. Marklund, G. T., Blomberg, L. G., Fälthammar, C.-G. & Lindqvist, P.-A. On intense shock-like electric fields associated with black aurora. *Geophys. Res. Lett.* **21**, 1859–1862 (1994).
8. Marklund, G. T., Blomberg, L. G., Fälthammar, C.-G., Lindqvist, P.-A. & Eliasson, L. On the occurrence and characteristics of intense low-altitude electric fields observed by Freja. *Ann. Geophys.* **13**, 704–712 (1995).
9. Marklund, G., Karlsson, T. & Clemmons, J. On low-altitude particle acceleration and intense electric fields and their relationship to black aurora. *J. Geophys. Res.* **102**, 17509–17522 (1997).
10. Carlson, C. W. *et al.* FAST observations in the downward auroral current region: energetic upgoing electron beams, parallel potential drops, and ion heating. *Geophys. Res. Lett.* **25**, 2017–2020 (1998).
11. Ergun, R. E. *et al.* FAST satellite observations of electric field structures in the auroral zone. *Geophys. Res. Lett.* **25**, 2025–2028 (1998).
12. Alfvén, H. On the theory of magnetic storms and aurorae. *Tellus* **10**, 104–116 (1958).
13. Borovsky, J. E. Auroral arc thicknesses as predicted by various theories. *J. Geophys. Res.* **98**, 6101–6138 (1993).
14. Carlqvist, P. & Boström, R. Space-charge regions above the aurora. *J. Geophys. Res.* **75**, 7140–7146 (1970).
15. Boström, R. Voltage drops along auroral magnetic field lines on small and large scale. *Phys. Space Plasmas* **15**, 37–42 (1998).
16. Temerin, M. & Carlson, C. W. Current–voltage relationship in the downward auroral current region. *Geophys. Res. Lett.* **25**, 2365–2368 (1998).
17. Rönmark, K. Electron acceleration in the auroral current circuit. *Geophys. Res. Lett.* **26**, 983–986 (1999).
18. Wygant, J. R. *et al.* Polar spacecraft based comparisons of intense electric fields and Poynting flux near and within the plasma sheet tail lobe boundary to UVI images: An energy source for the aurora. *J. Geophys. Res.* **105**, 18675–18692 (2000).
19. Kelling, A. *et al.* Properties of large electric fields in the plasma sheet at 4–7 R_E measured with Polar. *J. Geophys. Res.* **106**, 5779–5798 (2001).
20. Escoubet, C. P., Russel, C. T. & Schmidt, R. (eds) The Cluster and Phoenix Missions. *Space Sci. Rev.* **79** (1/2), 1–658 (1997).
21. Issue on Cluster first results. *Ann. Geophys.* **19** (October/November 2001).
22. Karlsson, T. & Marklund, G. Simulations of effects of small-scale auroral current closure in the return current region. *Phys. Space Plasmas* **15**, 401–406 (1998).
23. Vasyliunas, V. M. in *Particles and Fields in the Magnetosphere* (ed. McCormac, B.) 60–71 (Reidel, Dordrecht, 1969).
24. Lysak, R. L. Electrodynamic coupling of the magnetosphere and ionosphere. *Space Sci. Rev.* **52**, 33–87 (1990).

Acknowledgements

We are indebted to H. Frey for providing the IMAGE data for the event, and G. Gustavsson who was the Principal Investigator for the Electric Field and Wave (EFW) instrument until 2000.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.T.M. (e-mail: marklund@plasma.kth.se).

A one-dimensional chain state of vortex matter

Alexander Grigorenko*, Simon Bending*, Tsuyoshi Tamegai†, Shuichi Ooi‡ & Mohamed Henini‡

* Department of Physics, University of Bath, Claverton Down, Bath, BA2 7AY, UK

† Department of Applied Physics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-8627, Japan, and CREST, Japan Science and Technology Corporation (JST), Japan

‡ Department of Physics, University of Nottingham, Nottingham, NG7 2RD, UK

Magnetic flux penetrates isotropic type II superconductors in flux-quantized vortices, which arrange themselves into a lattice structure that is independent of the direction of the applied field¹. In extremely anisotropic high-transition-temperature (high- T_c) superconductors, a lattice of stacks of circular ‘pancake’ vortices forms when a magnetic field is applied perpendicular to the copper oxide layers, while an orthogonal elongated lattice of elliptical Josephson vortices forms when the applied field is parallel to the layers^{2–5}. Here we report that when a tilted magnetic field is applied to single crystals of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$, these lattices can interact to form a new state of vortex matter in which all stacks of pancake vortices intersect the Josephson vortices. The sublattice of Josephson vortices can therefore be used to manipulate the sublattice of pancake vortices. This result explains the suppression of irreversible magnetization by in-plane fields as seen in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ crystals, a hitherto mysterious observation⁶. The ability to manipulate sublattices could be important for flux-logic devices, where a ‘bit’ might be represented by a pancake vortex stack, and the problem of vortex positioning is overcome through sublattice interactions. This also enables the development of flux transducers and amplifiers, considerably broadening the scope for applications of anisotropic high- T_c superconductors.

The structure and organization of superconducting vortices directly reflects the crystalline anisotropy of the host material (Fig. 1A–D). A consequence of the extremely high anisotropy exhibited by $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ (BSCCO) single crystals is that tilted vortices spontaneously decompose into interacting ‘crossing lattices’ of pancake vortex (PV) and Josephson vortex (JV) stacks^{3–5} (Fig. 1E) whose phases remain largely unexplored. We show here the formation of a one-dimensional (1D) chain state of vortex matter in this regime, when all pancake vortex stacks become trapped on underlying stacks of Josephson vortices (Fig. 1F). This phase is bounded by transitions to a composite lattice^{7,8} of chains in a hexagonal lattice matrix at low tilt angles, and by a ‘sublimed’ chain state at high tilt angles.

We have used high-resolution scanning Hall probe microscopy⁹ to explore the crossing-lattices phase diagram in BSCCO single crystals under independently applied c -axis (H_c) and in-plane ($H_{||}$) magnetic fields. Magnetic images are generated parallel to the crystallographic a – b surface of freshly cleaved crystals with a lateral resolution of a few hundred nanometres, and the Hall probe typically detects the top 300 pancake vortices in a stack whose flux threads it along the c axis. The location of stacks of JVs is inferred from the fact that they become ‘decorated’ by PVs owing to their mutual attraction, in analogy with the well known Bitter decoration technique with fine ferromagnetic particles.

Figure 2A shows a typical set of images, obtained from scanning Hall probe microscopy, of the magnetic induction just above the face of a BSCCO single crystal as the angle between the applied field and the c axis was progressively increased. With a 12-Oe field parallel to the c axis at 81 K (Fig. 2A, a), a well-ordered hexagonal lattice of PV stacks was observed. However, as the in-plane field

component ($H_{||}$) was increased from zero, ‘chains’ of PV stacks formed, which separated domains of approximately hexagonal PV lattice (Fig. 2A, b and c). As seen in earlier Bitter decoration experiments^{7,8}, the chains were aligned along the direction of the in-plane field, and the vortex separation within the chains was smaller than that of the Abrikosov lattice. As H_c is decreased below the ‘ordering’ field, when intervortex interactions become comparable to pinning forces¹⁰, quenched disorder in the BSCCO crystal destroys the six-fold symmetry of the Abrikosov domains (Fig. 2A, d). As H_c is decreased still further, we observe a phase transition to another ordered 1D vortex chain state when the domains of Abrikosov lattice vanish entirely (Fig. 2A, e and f).

The crossing-lattices picture^{3–5} explains the complex behaviour illustrated in Fig. 2, where chain formation arises owing to the

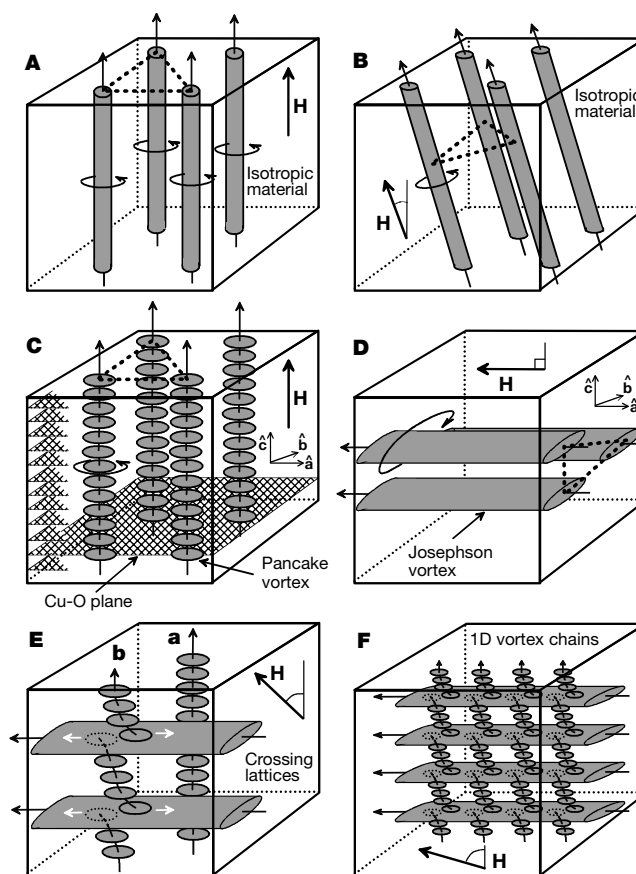


Figure 1 Sketches of vortex structures in isotropic and layered superconductors. **A, B**, The situation in an isotropic superconductor where the repulsion between vortices leads to the formation of an ordered hexagonal lattice independent of the tilt angle. Curved arrows indicate circulating supercurrents around the normal vortex core. **C**, Hexagonal ordering of the vortex lattice in layered superconductors with the magnetic field applied along the high-symmetry c axis. In this case, vortices are formed of vertical stacks of 2D pancake vortices situated in the Cu–O planes where superconductivity resides. **D**, With the magnetic field parallel to the layers, crystalline anisotropy leads to the formation of elliptical Josephson vortices whose ‘cores’ reside in the spaces between Cu–O planes, and whose circulating currents derive partly from strong supercurrents within the Cu–O planes and partly from weak Josephson coupling between them. Josephson vortices order into a highly elongated rhombic lattice. **E**, For a broad range of intermediate angles, tilted vortices spontaneously decompose into coexisting orthogonal PV and JV ‘crossing lattices’. **a**, If stacks of PVs and JVs do not intersect, no direct interaction occurs. **b**, Where a PV stack intersects a JV stack, small PV displacements (indicated by white arrows) driven by the underlying JV supercurrents lead to an attractive interaction. **F**, Zoomed-out view of the 1D vortex chain state when all PV stacks become trapped on vertical stacks of JVs.

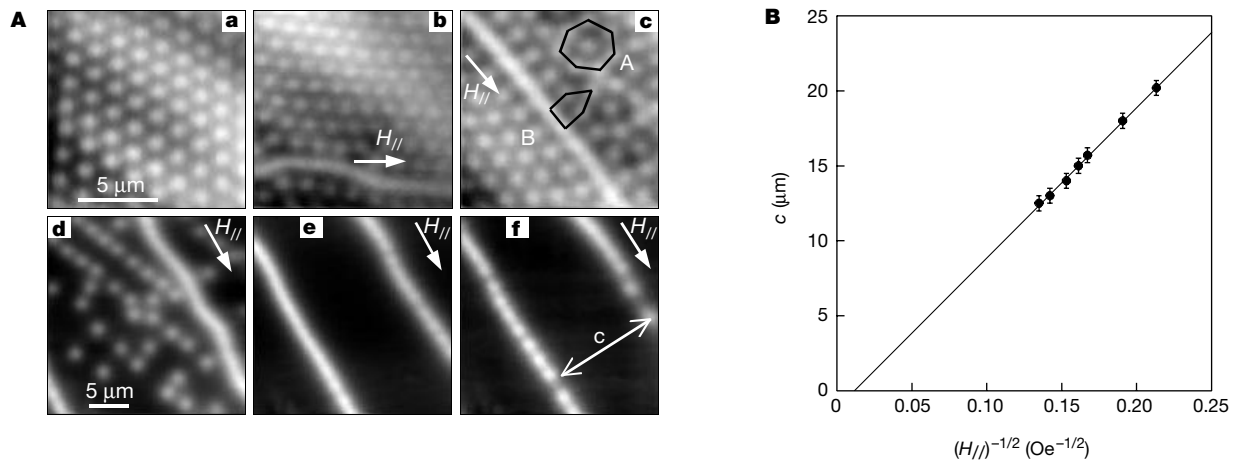


Figure 2 Images of stacks of pancake vortices in a BSCCO single crystal. The depth to which the Hall sensor probes the PV stacks in the as-grown crystals ($T_c = 90$ K) is roughly the in-plane magnetic penetration depth (typically the top 300 PVs in a stack). **A, a**, The hexagonal Abrikosov lattice of PV stacks obtained at $H_c = 12$ Oe ($H_{||} = 0$) at 81 K. (Greyscale spans 2.5 G.) The composite vortex lattice state is shown in **b** for $H_c = 14$ Oe and $H_{||} = 32$ Oe at 81 K (greyscale, 1.9 G), and for a different orientation of in-plane field (greyscale, 1.9 G) in **c** for $H_c = 10$ Oe and $H_{||} = 32$ Oe at 77 K. Arrows indicate the direction of the in-plane field. The presence of the incommensurate chains frustrates the surrounding hexagonal vortex lattice domains, giving rise, for example, to the seven-fold and five-fold rings indicated by A and B, respectively. **d**, The composite lattice state below the 'ordering' field at $H_c = 2$ Oe and $H_{||} = 27$ Oe at 81 K (greyscale,

2.9 G). The 1D vortex chain state is shown at **e**, $H_c = 1.2$ Oe and $H_{||} = 40$ Oe at 81 K (greyscale, 2.5 G), and with a lower density of PVs at **f**, $H_c = 0.5$ Oe and $H_{||} = 35$ Oe at 81 K. The chains of PV stacks in **e** and **f** are up to 50 in-plane penetration depths apart, and truly non-interacting. Comparing the PV–JV attraction and intra-chain PV–PV repulsion for the limit of non-overlapping JV cores⁵, we find that the isolated chain state is stable when $H_c^2/H_{||} > \Phi_0/(\sqrt{3}\gamma\lambda_{ab}^2\ln^2(c/c_0))$ where λ_{ab} is the in-plane penetration depth, $c_0 = A\lambda_{ab}^2/(\gamma s)$, A is a constant of order unity and s is the Cu–O plane separation. Experimentally we find that this ratio depends only weakly upon $H_{||}$, and is about 0.04 Oe at 77 K for our as-grown sample. **B**, Plot illustrating the linear dependence of the chain separation, c (as indicated in **A, f**), on $(H_{||})^{-1/2}$ at 81 K.

attraction of PVs to underlying stacks of JVs in the rhombic JV lattice, with an expected lateral separation of $c = [\sqrt{3}\gamma\Phi_0/(2H_{||})]^{1/2}$ (where γ is the anisotropy parameter and Φ_0 is the superconducting flux quantum). The measured chain separation, c , at 81 K is plotted as a function of $(H_{||})^{-1/2}$ in Fig. 2B, and the linear dependence

allows us to evaluate directly the anisotropy parameter $\gamma = 580 \pm 20$, in reasonable agreement with other estimates¹¹. PV stacks can always lower their energy by moving onto JV stacks provided that repulsive interactions with nearby pancake vortices do not become prohibitively large. Consequently the formation of

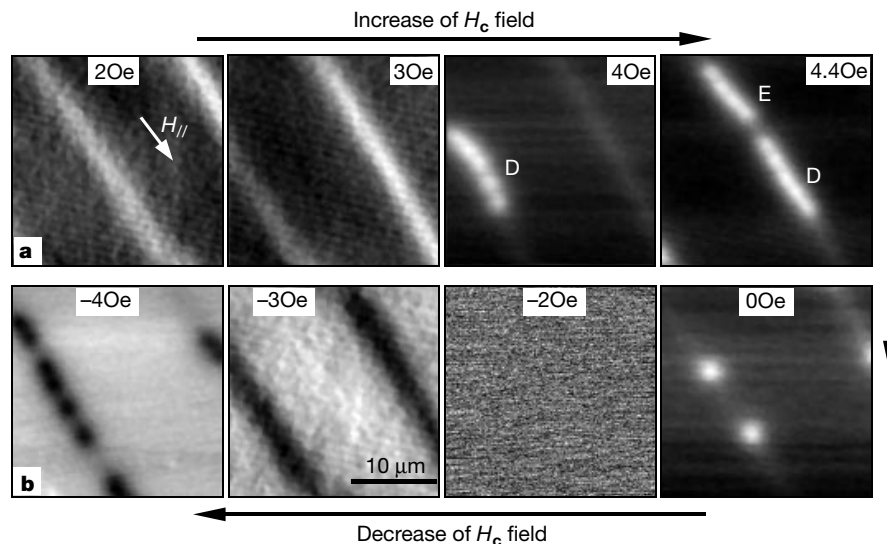


Figure 3 Images of pancake vortices in a BSCCO single crystal. This series of movie images shows the 'crystallization' and subsequent 'sublimation' of pancake vortex stacks after zero-field cooling to 83 K as the c -axis magnetic field is varied frame by frame, keeping the in-plane field fixed at $H_{||} = 38$ Oe. Each image takes ~ 12 s to acquire. (Entire movies and additional data are available at <http://www.bath.ac.uk/~pyssb/>.) **a**, H_c increasing. The 'sublimed' state is shown at $H_c = 2$ Oe (greyscale, 0.35 G) and 3 Oe (greyscale, 0.4 G), as well as the partially 'crystallized' state at $H_c = 4$ Oe (greyscale, 2.3 G) and 4.4 Oe (greyscale, 2.3 G). The coexistence of adjacent low flux density and high flux density regions seems to indicate that some form of phase separation has taken place. **b**, H_c decreasing. Shown are 'resublimation' of the crystallites as the field was

reduced to $H_c = 0.0$ Oe (greyscale, 1.7 G), and the total expulsion of PVs at $H_c = -2.0$ Oe (greyscale, 0.14 G), as well as the sublimed state in the reverse direction at $H_c = -3$ Oe (greyscale, 0.3 G) and its 'crystallization' at $H_c = -4$ Oe (greyscale, 2.1 G). Note that the reversal of sign of the 'sublimed' stripes as H_c is reversed at fixed in-plane field indicates that we are detecting pancake vortices and not some component of the field from the Josephson vortices. We speculate that the 'sublimed' state is one where pancake vortex stacks decompose into tilted vortices composed of a staircase of individual pancakes, or short PV segments, linked by sections of JV. We expect these objects to exhibit very large thermal fluctuations at the high temperatures used in these experiments.

the 1D chain state (Fig. 1F) may be anticipated for small values of H_c when the spacing between PV stacks along the chain can be arbitrarily large.

The 1D vortex chain state is bounded by a transition to a composite lattice composed of chains embedded in a hexagonal lattice 'matrix' at low tilt angles (Fig. 2A, b and c). But at high tilt angles, the phase again becomes unstable, and a transition to a state composed of 1D chains of 'sublimed' PV stacks is observed. Figure 3 shows a series of images from a movie of the sublimation process as the phase boundary is crossed in both directions at 83 K by cycling H_c slowly at fixed $H_{||} = 38$ Oe. At very small c -axis fields, the 1D vortex chain is only visible as a structureless 'sublimed' stripe of very low flux density ($H_c = 2$ Oe and 3 Oe). The magnetic induction (hence the PV density) associated with the 'sublimed' stripe grows with increasing H_c , implying that PVs preferentially penetrate and travel along JV stacks. But as H_c is increased further, 'crystallites' of well resolved PV stacks, with ten times higher flux density, nucleate and grow ($H_c = 4$ Oe and 4.4 Oe). If H_c is reduced again, these crystallites 'resublime' ($H_c = 0$ Oe), and eventually all PVs leave the sample ($H_c = -2$ Oe). If the field is decreased further, the sublimed stripes reverse sign ($H_c = -3$ Oe), and the nucleation of 'black' crystallites occurs ($H_c = -4$ Oe). Sublimation processes of this type have not (to our knowledge) been previously considered, although the first-order melting of vortex crystals in BSCCO as the magnetic field is increased has been widely studied^{12,13}. A second re-entrant melting phenomenon has also been theoretically predicted as the field is reduced in the solid phase; this transition is driven, not by an increase in entropy as for first-order melting, but by the exponential 'softening' of the crystal shear modulus at very low fields¹⁴. In contrast, we speculate that the transition to the 'sublimed' state corresponds to the shearing of PV stacks into tilted vortices composed of a staircase of isolated pancakes, or small pancake segments, linked by sections of Josephson vortex. Our experimentally established boundaries for the 1D vortex chain state are summarized in the experimental phase diagram for BSCCO single crystals under tilted magnetic fields shown in Fig. 4.

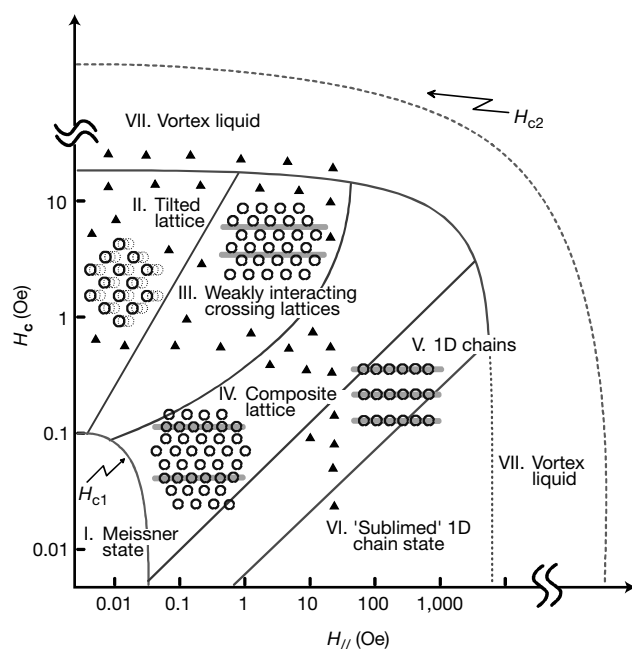


Figure 4 Experimental phase diagram for vortex matter in BSCCO single crystals. Our experimental data (filled triangles) have been used to map out a phase diagram for the different observed states of vortex matter in the H_c - $H_{||}$ domain for the temperature range where this study was performed (77–88 K).

For small H_c , we find that the underlying JV stacks represent robust 1D traps for PV stacks over a broad studied temperature range (77–88 K), while the repulsion between pancakes in the same Cu–O plane prevents PV stacks from passing one another along the length of the chain. As a consequence, the dynamic properties of the 1D vortex chain state appear to be particularly rich, as illustrated by the following two complementary experiments where one magnetic field component is held constant and the other varied.

Figure 5a shows images from a movie, illustrating how a trapped chain of PVs can be coherently dragged along by a JV stack as $H_{||}$ is reduced at fixed $H_c (= 0.5$ Oe) at 83 K. During its movement from the lower left to the upper right corner, the JV stack (J) is able to depin and drag with it two isolated PV stacks (A, B), which were originally pinned at defects on the right-hand side of the image. Pancake vortex displacements can be uniquely and reversibly determined by adjusting the applied field component $H_{||}$; we consider that this vortex 'pump'-like action may find applications in flux-logic devices where a 'bit' is defined by the presence (or absence) of a single PV stack. A simple binary switch could, for example, be realized by using an in-plane field to manipulate a single PV stack underneath a flux transducer (for example, a microscopic induction coil or Hall probe) whose voltage output provides a read-out signal. The small pancake vortex diameters (~400 nm) potentially allow very-large-scale integration of such devices with high operation speeds. In-plane magnetic fields could also be used to indirectly compress PV stacks, yielding an active PV

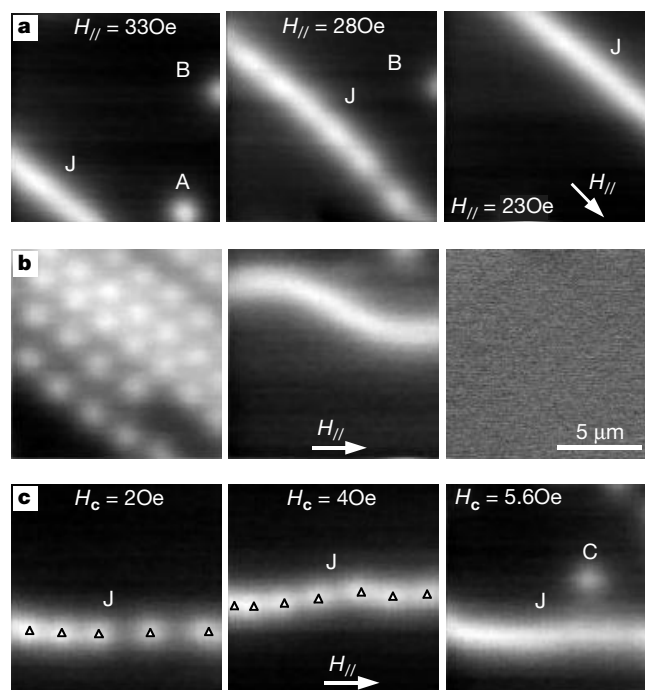


Figure 5 Images from a movie of the motion of pancake vortex stacks as the magnetic field is varied. **a**, The vortex 'pump'-like action achieved by varying $H_{||}$ with fixed $H_c (= 0.5$ Oe) at 83 K: left, $H_{||} = 33$ Oe (greyscale, 2.9 G); middle, $H_{||} = 28$ Oe (greyscale, 3.2 G); and right, $H_{||} = 23$ Oe (greyscale, 3.4 G). **b**, The demagnetization procedure that uses this vortex 'pump' principle: left, the remnant state after application and removal of $H_c = 4$ Oe at 85 K (greyscale, 4.3 G); middle, application of $H_{||} = 35$ Oe ($H_c = 0$) at 85 K (greyscale, 3.5 G); right, the demagnetized sample after $H_{||}$ is slowly reduced to zero (greyscale, 0.6 G). **c**, Images from a movie of the penetration of PVs along 1D chains as H_c is increased after field-cooling to 81 K at $H_c = 2$ Oe and (fixed) $H_{||} = 35$ Oe: left, $H_c = 2$ Oe (greyscale, 3.2 G); middle, $H_c = 4$ Oe (greyscale, 3.7 G); and right, $H_c = 5.6$ Oe (greyscale, 4.2 G). The PV density along the chain J (indicated by triangles) increases until at saturation the PV stack, C, moves into the domain between JV stacks.

flux amplifier or transducer. Figure 5b illustrates how the pump can be used to demagnetize a sample¹⁵. The left-hand frame shows the flux trapped in our sample after the application and subsequent removal of $H_c = 4$ Oe at 85 K ($H_{||} = 0$). An in-plane field, $H_{||} = 35$ Oe, was then applied, driving the sample into the vortex chain state (middle frame). Finally, the in-plane field $H_{||}$ was slowly removed and, as the JVs were swept out of the sample they dragged the PVs with them, leaving a demagnetized sample (right-hand frame).

Magnetic irreversibility in layered superconductors is known to be suppressed by the application of either a.c.^{6,15} or d.c. in-plane magnetic fields. The movie images shown in Fig. 5c yield insights into the latter situation, where H_c has been increased from 2 Oe at fixed $H_{||} = 35$ Oe at 81 K. We find that PVs preferentially penetrate along the 1D vortex chains (left hand and middle frames), and only propagate into the inter-chain spaces if the chain density becomes saturated (right-hand frame) as explained in Fig. 2 legend. It seems that the PV stack structure induced by the JV currents weakens the interaction with quenched sample disorder, and the mobility of PVs along the 1D chains is considerably higher than in the inter-chain spaces. Magnetization loops measured with the scanning Hall probe retracted from the sample surface confirm this conclusion, and reveal a marked suppression of the irreversibility in the presence of JVs ($H_{||} \neq 0$). PV penetration in this regime is known to be limited by electromagnet surface barriers¹⁶, which will be slightly lower where JVs intercept the edges, owing to the superposition of Meissner and JV currents. Hence PVs preferentially enter at the sample edges along JVs where surface barriers are lowest, and then show a much higher mobility along the 1D vortex chains. □

Received 2 May; accepted 19 October 2001.

1. Kleiner, W. H., Roth, L. M. & Autler, S. H. Bulk solution of Ginzburg-Landau equations for type II superconductors: upper critical field region. *Phys. Rev. A* **133**, 1226–1227 (1964).
2. Clem, J. R. Anisotropy and two-dimensional behaviour in the high-temperature superconductors. *Supercond. Sci. Technol.* **11**, 909–914 (1998).
3. Bulaeviski, L. N., Levdi, M. & Kogan, V. G. Vortices in layered superconductors with Josephson coupling. *Phys. Rev. B* **46**, 366–380 (1992).
4. Huse, D. A. Magnetic-flux patterns on the surface of a type-II superconductor. *Phys. Rev. B* **46**, 8621–8623 (1992).
5. Koshelev, A. E. Crossing lattices, vortex chains, and angular dependence of melting line in layered superconductors. *Phys. Rev. Lett.* **83**, 187–190 (1999).
6. Farrell, D. E. *et al.* Magnetization jumps and irreversibility in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. B* **53**, 11807–11816 (1996).
7. Bolle, C. A. *et al.* Observation of a commensurate array of flux chains in tilted flux lattice in Bi-Sr-Ca-Cu-O single crystals. *Phys. Rev. Lett.* **66**, 112–115 (1991).
8. Grigorieva, I. V., Steeds, J. W., Balakrishnan, G. & Paul, D. M. Vortex-chain state in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ —experimental evidence for coexistence of 2 vortex orientations. *Phys. Rev. B* **51**, 3765–3771 (1995).
9. Oral, A., Bending, S. J. & Henini, M. Real-time scanning Hall probe microscopy. *Appl. Phys. Lett.* **69**, 1324–1326 (1996).
10. Grier, D. G. *et al.* Translational and bond-orientational order in the vortex lattice of the high-T superconductor $\text{Bi}_{1.9}\text{Sr}_{1.1}\text{CaCu}_2\text{O}_{8+x}$. *Phys. Rev. Lett.* **66**, 2270–2273 (1991).
11. Martinez, J. C. *et al.* Magnetic anisotropy of a $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_x$ single crystal. *Phys. Rev. Lett.* **69**, 2276–2279 (1992).
12. Zeldov, E. *et al.* Thermodynamic observation of first-order vortex-lattice melting transition in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Nature* **375**, 373–376 (1995).
13. Soibel, A. *et al.* Imaging the vortex-lattice melting process in the presence of disorder. *Nature* **406**, 282–287 (2000).
14. Nelson, D. R. Vortex entanglement in high- T_c superconductors. *Phys. Rev. Lett.* **60**, 1973–1976 (1988).
15. Avraham, N. *et al.* Inverse melting of a vortex lattice. *Nature* **411**, 451–454 (2001).
16. Burlachkov, L. *et al.* Giant flux-creep through surface barriers and the irreversibility line in high temperature superconductors. *Phys. Rev. B* **50**, 16770–16773 (1994).

Acknowledgements

We thank M.J.W. Dodgson for discussions. This work was supported in the UK by the EPSRC and the MOD, in the USA by MARTECH, Florida State University, and in Japan by the Ministry of Education, Science, Sports and Culture.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.B. (e-mail: pyssb@bath.ac.uk).

High-temperature ultrafast polariton parametric amplification in semiconductor microcavities

M. Saba*, C. Ciuti*, J. Bloch†, V. Thierry-Mieg†, R. André‡, Le Si Dang‡, S. Kundermann*, A. Mura§, G. Bongiovanni§, J. L. Staehli* & B. Deveaud*

* Physics Department, Swiss Federal Institute of Technology Lausanne, PH-Ecublens, CH-1015 Lausanne-EPFL, Switzerland

† Centre National de la Recherche Scientifique, L2M-CNRS, 92225 Bagneux Cedex, France

‡ Laboratoire de Spectrométrie Physique, Université J. Fourier-Grenoble, F-38402 Saint Martin d'Hères Cedex, France

§ Dipartimento di Fisica and Istituto Nazionale di Fisica della Materia, Università degli Studi di Cagliari, I-09042 Monserrato, Italy

Cavity polaritons, the elementary optical excitations of semiconductor microcavities, may be understood as a superposition of excitons and cavity photons¹. Owing to their composite nature, these bosonic particles have a distinct optical response, at the same time very fast and highly nonlinear. Very efficient light amplification due to polariton–polariton parametric scattering has recently been reported in semiconductor microcavities at liquid-helium temperatures^{2–11}. Here we demonstrate polariton parametric amplification up to 120 K in GaAlAs-based microcavities and up to 220 K in CdTe-based microcavities. We show that the cut-off temperature for the amplification is ultimately determined by the binding energy of the exciton. A 5- μm -thick planar microcavity can amplify a weak light pulse more than 5,000 times. The effective gain coefficient of an equivalent homogeneous medium would be 10^7 cm^{-1} . The subpicosecond duration and high efficiency of the amplification could be exploited for high-repetition all-optical microscopic switches and amplifiers. 10^5 polaritons occupy the same quantum state during the amplification, realizing a dynamical condensate of strongly interacting bosons which can be studied at high temperature.

A semiconductor microcavity is a photonic structure designed to enhance light–matter interaction. The cavity photons are confined between two mirrors, and resonantly interact with the excitonic transition of a two-dimensional semiconductor quantum well (Fig. 1). In the strong-coupling regime, the normal modes of the system are the cavity polaritons, which are half-exciton, half-photon quasiparticles¹. The energies of the two polariton modes anticross when the energy difference between exciton and photon modes is varied (Fig. 1). The minimum polariton splitting measures the strength of the coupling. Owing to their excitonic content, polaritons are subject to Coulomb interaction and give rise to strong optical nonlinearities. At the same time, owing to the photon component, the curve of polariton energy dispersion versus in-plane wavevector is very steep, meaning that it is quite sensitive to the excitation angle (Fig. 1). A small number of states can be excited at a given angle, and quantum degeneracy (more than one polariton per state) is achieved with relatively low excitation densities. The bosonic statistics strongly enhances some particular scattering processes, provided that the density is kept below the critical value where the bosonic behaviour of excitons breaks down owing to the appearance of the fermionic nature of electrons and holes, which make up the excitons¹². The scattering from an incoherent exciton reservoir into polaritons can be stimulated by the occupation of the final state¹³ and give rise to a polariton laser¹⁴. If, as in our case, coherent polaritons are injected into the cavity, the phase-coherent final-state stimulation (termed parametric amplification) can be much more efficient than in the incoherent case^{2–11}.

To investigate the coherent regime, polaritons are resonantly excited by a pump laser pulse on the lower-polariton dispersion curve² (Fig. 1). The excitation angle is chosen to allow energy and momentum conservation for the scattering between two polaritons, one into the lowest-energy state (with in-plane wavevector $k = 0$) and the other into a higher-energy state ($2E_k = E_0 + E_{2k}$ has to be fulfilled, where E_k is the energy of a pump polariton injected with an in-plane wavevector k). A weak probe beam (the 'signal'), injected perpendicular to the cavity ($\theta = 0$), stimulates the scattering into the lowest-energy state at the bottom of the dispersion curve ($k = 0$). This non-degenerate wave-mixing process can also be seen as optical parametric amplification. The exciton content makes the polariton–polariton interaction very efficient⁵, and gains of up to 100 on the probe beam have been reported at liquid-helium temperatures². The spontaneous relaxation of a few pump polaritons into the lowest-energy state can also trigger parametric scattering under continuous-wave optical pumping and without an external probe^{7–11}. In order to exploit this nonlinear phenomenon in practical devices, one of the challenges is to demonstrate the polariton parametric process at high temperatures.

We show here that engineering the sample structure in order to maximize the exciton–photon coupling greatly increases the parametric polariton amplification, giving optical gains up to 5,000. The amplification survives at high temperatures, up to 220 K in the CdTe microcavity, approaching room temperature and allowing the design of several possible devices. We find that in the structures

we considered, the cut-off temperature of the parametric gain scales linearly with the binding energy of the exciton.

The polariton splitting in microcavities is proportional to the square root of the oscillator strength of the excitonic transition, and can be increased by inserting a large number of quantum wells into the cavity. Here we consider GaAlAs-based and CdTe-based microcavities grown by molecular-beam epitaxy. The 3λ CdTe-based cavity contains six stacks of four quantum wells, at the antinodes of the electric field, giving a 25-meV low-temperature polariton splitting. In the GaAlAs-based microcavities, a few quantum-well stacks are also placed in the dielectric mirrors (distributed Bragg reflectors), at positions corresponding to the maxima of light intensity¹⁵ (Fig. 1). With this arrangement, the exciton also couples to the light field penetrating the mirrors, giving large polariton splittings—15.3 meV (12 quantum wells) and 20 meV (36 wells)—in spite of the lower excitonic oscillator strength of the GaAlAs-based structures compared to CdTe-based structures. In all three samples, the polariton anticrossing is observed up to room temperature.

The pump–probe measurements presented here are taken in reflection geometry: the sample is illuminated by transform-limited 250-fs-long pulses from a mode-locked Ti:sapphire laser with an 80-MHz repetition rate. A mechanical goniometer allows us to vary the pump angle while keeping constant the pump–probe delay; the weak probe beam hits the sample nearly at normal incidence. To avoid a non-uniform excitation, the centre of the probe spot on the sample is spatially selected by a pin-hole.

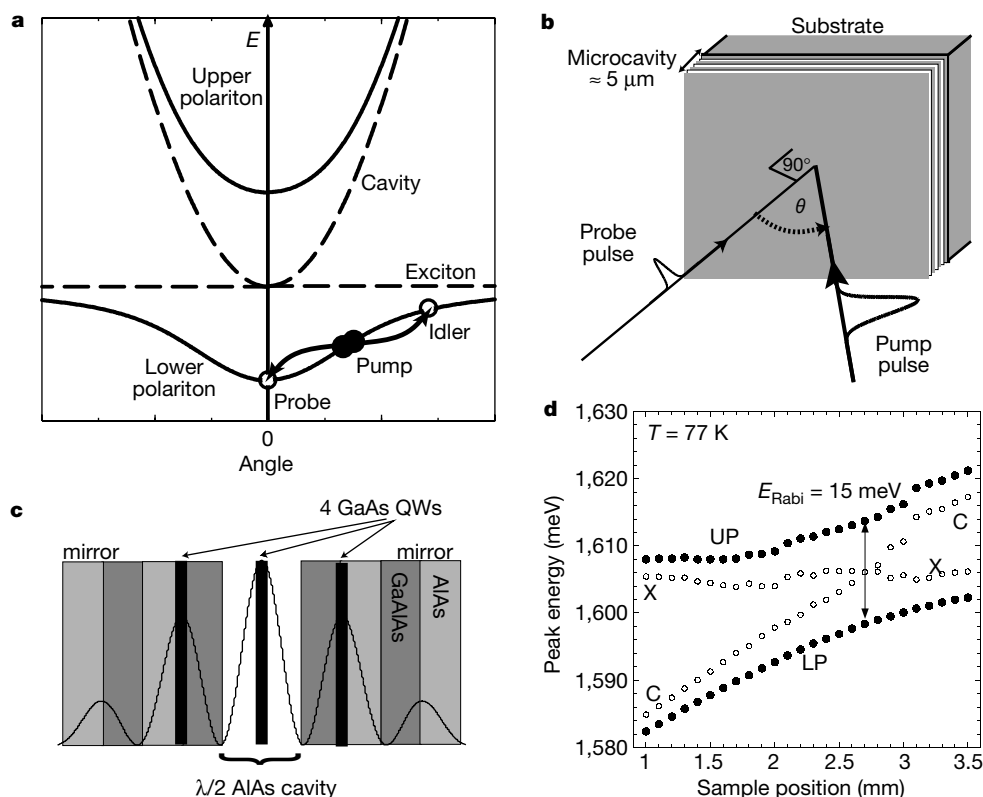


Figure 1 Principle of the pump–probe optical set-up and sample characteristics. **a**, Sketch of the polariton energy E versus the incidence angle θ of the exciting light. The corresponding in-plane wavevector k of the excited polariton is given by $k = (\omega/c) \sin \theta$, where ω is the light frequency. The parametric polariton–polariton scattering process is also depicted. **b**, Angle-resolved pump–probe configuration. **c**, GaAlAs-based sample structure, with three stacks of four quantum wells (QWs). The 7-nm-wide GaAs wells are separated by 3-nm-wide AlAs barriers. The Bragg mirrors consist of alternating $\text{Ga}_{0.8}\text{Al}_{0.2}\text{As}$ /AlAs $\lambda/4$ layers. The solid line depicts the qualitative behaviour of the electric field intensity in the cavity. Another sample was prepared with nine stacks of wells¹⁴ (36 wells). In the CdTe-based sample, six stacks of four wells each are placed at the six

antinodes of the electric field inside a 3λ cavity. The 24 5-nm-wide CdTe wells are separated by 6-nm-wide $\text{CdMg}_{0.6}\text{Te}$ barriers. Alternating $\lambda/4$ layers of $\text{CdMg}_{0.6}\text{Te}$ and $\text{CdMn}_{0.25}\text{Te}$ act as cavity mirrors. **d**, The cavities are wedge-shaped, so that by varying the position on the sample the cavity energy also varies. By tuning the cavity energy (G) with respect to the excitonic energy (X , which is instead almost constant), the anticrossing of the polariton energies is observed. The polariton energies (filled circles) for the GaAlAs sample are measured from the linear reflectivity spectra, while the exciton and the cavity energies (open circles) are extracted from the data with a two coupled oscillators model. LP, lower-polariton energy; UP, upper-polariton energy; E_{Rabi} , minimum polariton splitting.

The probe beam is greatly amplified (Fig. 2)—by a factor of up to 5,000 in one of our GaAlAs-based samples, a gain nearly two orders of magnitude higher than in previous reports². The strong line appearing in the gain spectra is narrower than the lower polariton, and lies between the energy of the unperturbed polariton and that of the cavity mode. Because the interaction between co-polarized polaritons (through their excitonic components) turns out to be repulsive, the whole polariton branch is shifted towards higher energies, and the gain spectrum peaks on the high-energy tail of the unperturbed polariton absorption line⁵. The maximum gain occurs at the angle corresponding to the inflection point of the dispersion curve of the lower polariton, where the energy and momentum conservation for the parametric scattering is most easily satisfied. The width of the angular resonance depends on the broadness of the energy distribution of pump polaritons, and is therefore related to the polariton spectral linewidth. The curve of gain versus pump power shows a sharp threshold, typical of stimulated scattering. Close to threshold, the probe gain can be increased by a factor of 10 by increasing the pump power by less than 50%, a highly desirable feature for an optical switch device.

Using the parameters of the experiment, we calculate that each probe pulse populates each state with approximately 100 polaritons (details on how to count the available polariton states are given in

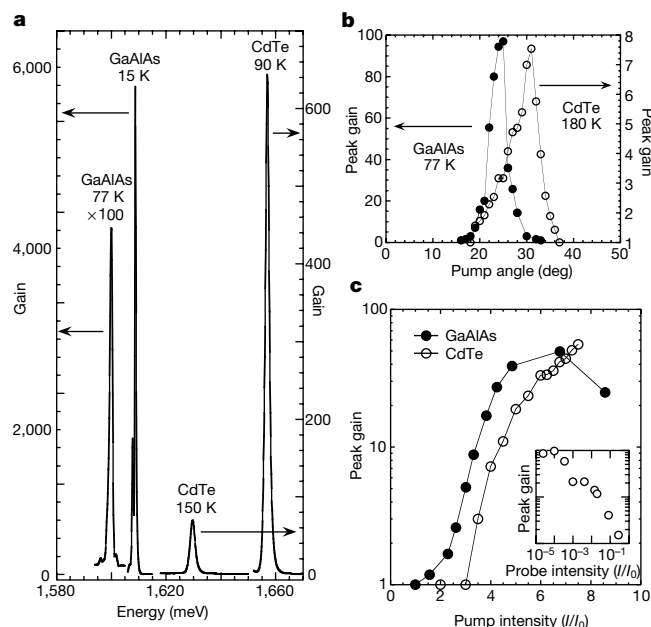


Figure 2 Main features of parametric gain: high efficiency, spectrally narrow lines, angular resonance and pump-power threshold. **a**, Gain spectra for different temperatures in the GaAlAs-based (12 wells) and CdTe-based samples. When the temperature increases, the polariton mode gets broader because of the thermal dephasing and consequently the gain peak gets smoother, broader and less intense. The energy of the polariton, and consequently that of the parametric gain, varies with temperature, as it follows the bandgap thermal redshift. The maximum gain occurs when the cavity mode energy is lower than the exciton energy by about half the polariton splitting. **b**, Angular resonance: in the CdTe sample, the maximum gain occurs at a higher angle than in the GaAlAs sample (the polariton splitting is larger) and the angular resonance is broader owing to the broader polariton line. **c**, The power dependence of peak gain (the highest value extracted from the spectra at the resonance angle) shows a threshold and then saturates at high powers. $I_0 = 10^{13}$ photons per cm^2 per pulse. The temperature is 77 K for the GaAlAs sample, 150 K for the CdTe sample. The threshold is reached at lower fluences in the GaAlAs sample owing to the lower number of wells inside the cavity and the smaller excitonic oscillator strength. Inset, the gain (normalized to its maximum value) is reduced by raising probe power (data are for the CdTe sample at $T = 90$ K; for other samples and temperatures, similar behaviour is found).

ref. 8). The parameters used for this estimate are the probe-beam angle spread (0.5°) and the probe photon density injected into the cavity (2×10^6 photons per cm^2 per pulse). When the gain is 5,000, the light intensity emitted in the probe direction after the parametric amplification is 2,500 times that absorbed from the probe: this is because the gain peak is approximately twice as narrow as the polariton. The resulting occupancy of the lowest-energy state is of the order of 10^5 polaritons per state. Even taking only 20% of the exciton fraction for each polariton, more than 10,000 excitons are in the same state. The device thickness along the growth direction is about $5 \mu\text{m}$, so the amplification would correspond, if directly expressed as an effective material gain, to an impressive coefficient of 10^7 cm^{-1} .

The parametric gain in the cavities featuring large polariton splittings is robust enough to survive up to 125 K in the GaAlAs samples and up to 220 K in the CdTe sample (Fig. 3). We note that

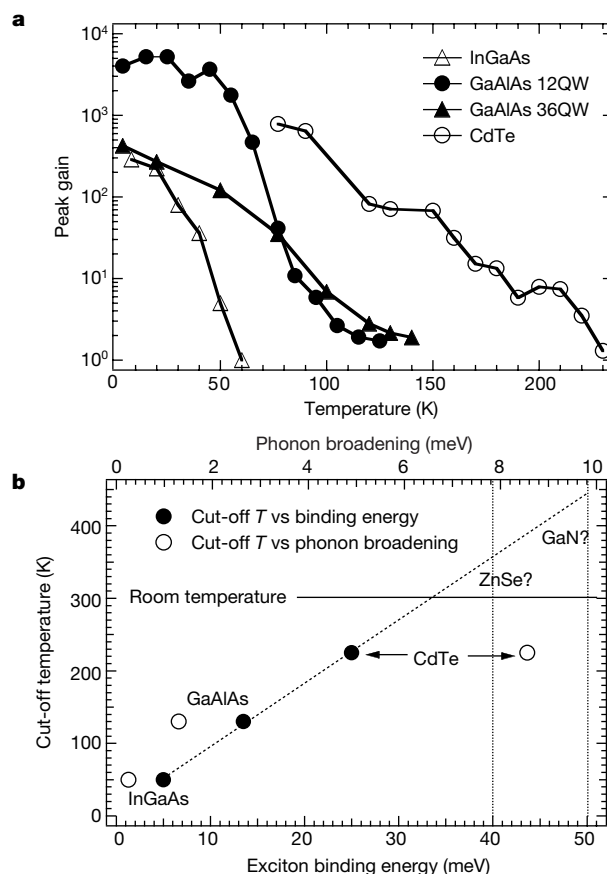


Figure 3 Temperature dependence of parametric gain. **a**, Peak gain versus temperature for the different samples. The CdTe cavity is designed to match the exciton energy at high temperatures; below 77 K, exciton–photon resonance is no longer achieved. The measurements are taken for the same pump fluence for both GaAlAs samples, but owing to the greater threshold for the 36-well sample, at low temperatures the gain is lower than in the 12-well sample (the pump fluence needed to saturate the gain in the 36-well sample could not be achieved with our laser). **b**, Cut-off temperatures for the gain in different materials as a function of the exciton binding energy (lower x axis, filled circles). The broken line is a linear fit to the data. The inverse logarithmic slope of the last points on each curve has the same dependence on the binding energy. The vertical dotted lines mark the typical exciton binding energies in ZnSe- and GaN-based quantum wells. In contrast, the cut-off temperature seems not to scale with the line broadening owing to the exciton–phonon coupling (upper x axis, open circles). The phonon broadening at the cut-off temperature is the product of the number of phonons in the lattice (determined only by the temperature and the phonon energy, 36 meV in GaAs, 21 meV in CdTe) times the exciton–phonon coupling coefficient (13.8 meV in GaAs, 15 meV in CdTe).

the gains of the two GaAlAs samples (one with 12 wells, and one with 36) show a very similar cut-off temperature, although the two samples have different polariton splittings. (The cut-off temperature is defined as the temperature at which the gain falls to 1, the highest possible operation temperature for the device.) This feature indicates that beyond a certain point the cut-off temperature does not rise any more with the polariton splitting, and is limited by some intrinsic parameter of the material. The CdTe sample, having a polariton splitting only 25% larger than that of the GaAlAs samples, sustains the parametric amplification up to almost twice the temperature. What is very different in the two materials is the exciton binding energy—that is, the energy separating the quasi-bosonic bound exciton states from the fermionic continuum of unbound electron–hole pairs. In the 7-nm-wide GaAs quantum wells, the exciton binding energy is about 13.5 meV, whereas in the CdTe wells it is about 25 meV. As a check of the binding-energy dependence, we repeated the experiment on an InGaAs-based microcavity with a 5-meV exciton binding energy, and found that it only shows gain up to 50 K. The dependence of the cut-off temperature on the exciton binding energy appears to be linear. This indicates that the cut-off temperature is related to the dissociation of polaritons or excitons into free carriers. In the linear regime this process does not significantly alter the optical response, as the polariton peaks are still well resolved at the cut-off temperature and the polariton splitting is reduced by only 10% with respect to the low-temperature value. The presence of fermionic free carriers has instead a marked effect on the nonlinear regime, as the parametric amplification critically depends on a limitation on the occupancy per state and on the degree of coherence of the polaritons⁵.

We note that the cut-off temperature scales linearly with the exciton binding energy in GaAs and CdTe microcavities, whereas it does not seem to depend on the phonon parameters (Fig. 3). The optical phonon energy and the matrix element of the exciton–phonon interaction strongly depend on the material parameters, and are expected to contribute to the efficiency of thermal dissociation of polaritons. Further investigations of other structures with larger exciton binding energies will help to understand the exact effect of the exciton–phonon interaction on the parametric gain. Suitable materials for such studies are some II–VI compounds, such as ZnSe, which have exciton binding energies as high as 40 meV (ref. 16). Furthermore, the exciton binding energy (in sufficiently

narrow wells) can be made larger than the optical phonon energy (as in CdTe), thus greatly suppressing the thermal ionization process. We expect GaN-based semiconductor compounds to attract attention in the future, owing to their very high oscillator strengths^{17,18}. Efforts at present are concentrated on the growth technology of nitride compounds in order to improve the sample quality and achieve strong coupling in microcavities. The strong-coupling regime at room temperature has already been demonstrated in organic microcavities^{19,20}, with large polariton splittings of around 100 meV. Even if the coherence time of optical excitations in such compounds is very short, the low material costs and simple sample preparation will encourage studies of ultrafast polaritonic nonlinearities.

The time dynamics of the parametric amplification can be explored by delaying the probe pulse with respect to the pump. The amplification lasts as long as pump and probe polariton waves are simultaneously inside the cavity, leading to ultrafast dynamics (Fig. 4). Moreover, the polaritons injected by the pump escape from the cavity within a few picoseconds after the excitation, so that a device based on polariton parametric amplification could have a very fast recovery time, with a physical limit for the repetition rate in the terahertz range. Further, because the light of the amplified probe pulse comes from a collective coherent polariton state, control of the polariton wavefunction inside the cavity by means of optical pulses could result in complete and ultrafast control of the amplitude, phase and spectral shape of the emitted light.

The massive occupation of the lowest-energy polariton state can be also seen as a dynamical condensation of polaritons. In comparison to the Bose condensation of atoms, polariton condensation occurs for strongly interacting particles⁴, at higher densities (the ratio between the particle spacing and their radius is of the order of 10 for polaritons and about 1,000 for atoms), with a much stronger coupling with the environment (a typical loss rate in microcavities is 10^{12} s^{-1} , compared to 10^5 s^{-1} for atomic systems): most notably, polariton condensation occurs at temperatures of the order of 100 K instead of 10^{-8} K . Semiconductor microcavities therefore offer the possibility of studying at 100 K the non-classical states of matter realized in dynamical condensates. □

Received 12 June; accepted 8 October 2001.

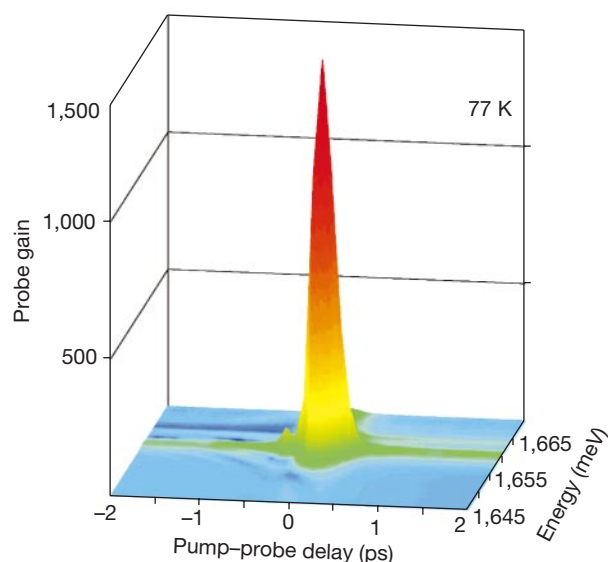


Figure 4 The probe reflectivity spectrum of the CdTe sample as a function of the pump–probe delay, showing the ultrafast dynamics of the parametric gain.

- Weisbuch, C., Nishioka, M., Ishikawa, A. & Arakawa, Y. Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Phys. Rev. Lett.* **69**, 3314–3317 (1992).
- Savvidis, P. G. *et al.* Angle-resonant stimulated polariton amplifier. *Phys. Rev. Lett.* **84**, 1547–1551 (2000).
- Ehrenstein, D. Amplifier from half-breed particles. *Phys. Rev. Focus* **5**, story 6 (2000).
- Yamamoto, Y. Half-matter, half-light amplifier. *Nature* **405**, 629–630 (2000).
- Ciuti, C., Schwendemann, P., Deveaud, B. & Quattropani, A. Theory of the angle-resonant polariton amplifier. *Phys. Rev. B* **62**, R4825–R4828 (2000).
- Dasbach, G., Baars, T., Bayer, M., Larionov, A. & Forchel, A. Coherent and incoherent polaritonic gain in a planar semiconductor microcavity. *Phys. Rev. B* **62**, 13076–13083 (2000).
- Houdré, R., Weisbuch, C., Stanley, R. P., Oesterle, U. & Ilegems, M. Nonlinear emission of semiconductor microcavities in the strong coupling regime. *Phys. Rev. Lett.* **85**, 2793–2796 (2000).
- Stevenson, R. M. *et al.* Continuous wave observation of massive polariton redistribution by stimulated scattering in semiconductor microcavities. *Phys. Rev. Lett.* **85**, 3680–3683 (2000).
- Savvidis, P. G. *et al.* Asymmetric angular emission in semiconductor microcavities. *Phys. Rev. B* **62**, R13278–R13281 (2000).
- Baumberg, J. J. *et al.* Parametric oscillation in a vertical microcavity: a polariton condensate or micro-optical parametric oscillation. *Phys. Rev. B* **62**, R16247–R16250 (2000).
- Ciuti, C., Schwendemann, P. & Quattropani, A. Parametric luminescence of microcavity polaritons. *Phys. Rev. B* **63**, 041303R–1–041303R–4 (2001).
- Imamoglu, A., Ram, R. J., Pau, S. & Yamamoto, Y. Nonequilibrium condensates and lasers without inversion: exciton-polariton lasers. *Phys. Rev. A* **53**, 4250–4253 (1996).
- Huang, R., Tassone, F. & Yamamoto, Y. Experimental evidence of stimulated scattering of excitons into microcavity polaritons. *Phys. Rev. B* **61**, R7854–R7857 (2000).
- Le Si Dang, Heger, D., André, R., Boeuf, F. & Romestain, R. Stimulation of polariton photoluminescence in semiconductor microcavity. *Phys. Rev. Lett.* **81**, 3920–3923 (1998).
- Bloch, J., Freixanet, T., Marzin, J. Y., Thierry-Mieg, V. & Paniel, R. Giant Rabi splitting in a microcavity containing distributed quantum wells. *Appl. Phys. Lett.* **73**, 1694–1696 (1998).
- Kelkar, P. V. *et al.* Stimulated emission, gain and coherent oscillations in II–VI semiconductor microcavities. *Phys. Rev. B* **56**, 7564–7573 (1997).
- Ponce, F. A. & Bour, D. P. Nitride-based semiconductors for blue and green light-emitting devices. *Nature* **386**, 351–359 (1997).
- Pearton, S. J., Zolper, J. C., Shul, R. J. & Ren, F. GaN: processing, defects, and devices. *J. Appl. Phys.* **86**, 1–78 (1999).

19. Lidzey, D. G. *et al.* Strong exciton–photon coupling in an organic semiconductor microcavity. *Nature* **395**, 53–56 (1998).
20. Lidzey, D. G., Bradley, D. D. C., Armitage, A., Walker, S. & Skolnick, M. S. Photon-mediated hybridisation of Frenkel excitons in organic semiconductor microcavities. *Science* **288**, 1620–1623 (2000).

Acknowledgements

We thank U. Oesterle and R. Houdré for growing the InGaAs sample, and J.D. Ganiere, G. Hayes, A. Quattropani, R. Romestain, P. Senellart and P. Schwendimann for suggestions. This work has been supported in part by the Fonds National Suisse de la Recherche Scientifique.

Correspondence and requests for materials should be addressed to M.S. (e-mail: Michele.Saba@epfl.ch).

Hierarchical self-assembly of metal nanostructures on diblock copolymer scaffolds

Ward A. Lopes^{*†} & Heinrich M. Jaeger^{*}

^{*} James Franck Institute and Department of Physics, The University of Chicago, Chicago, Illinois 60637, USA

Self-assembly is emerging as an elegant, ‘bottom-up’ method for fabricating nanostructured materials^{1–8}. This approach becomes particularly powerful when the ease and control offered by the self-assembly of organic components is combined with the electronic, magnetic or photonic properties of inorganic components^{2,5,9}. Here we demonstrate a versatile hierarchical approach for the assembly of organic–inorganic, copolymer–metal nanostructures in which one level of self-assembly guides the next. In a first step, ultrathin diblock copolymer films form a regular scaffold of highly anisotropic, stripe-like domains^{10–12}. During a second assembly step, differential wetting guides diffusing metal atoms to aggregate selectively along the scaffold, producing highly organized metal nanostructures. We find that, in contrast to the usual requirement of near-equilibrium conditions for ordering^{2,3,13}, the metal arranged on the copolymer scaffold produces the most highly ordered configurations when the system is far from equilibrium. We delineate two distinct assembly modes of the metal component—chains of separate nanoparticles and continuous wires—each characterized by different ordering kinetics and strikingly different current–voltage characteristics. These results therefore demonstrate the possibility of guided, large-scale assembly of laterally nanostructured systems.

Two important, competing, issues for the assembly of inorganic materials on organic scaffolds are high density and selectivity. In the case of scaffolds formed by copolymer domains (Fig. 1), preferential wetting of one of the copolymer blocks by metal will selectively aggregate the metal inside the corresponding domain^{14,15}. However, simple coalescence of dense metal nanoparticle aggregates into the overall shape given by the boundaries of the selected copolymer domain—similar to that observed for water droplets filling out a narrow hydrophilic strip³—has not been achieved so far. This is because of the large surface energies of metals, exceeding those of copolymers by orders of magnitude. Consequently, metal–metal bonds will overwhelm metal–polymer bonds and, except for very small metal concentrations, the final configuration will be a large, spherical metal aggregate that completely ignores the polymer

scaffold and mitigates selectivity. Thus, under equilibrium conditions even highly elongated, anisotropic polymer scaffolds may not be able to guide self-assembling metal particles into wire-like nanostructures. But here we show how non-equilibrium processes can induce, and stabilize, recognition of complex patterns provided by a nanoscale scaffold.

The particular, asymmetric polystyrene-block-poly(methyl-methacrylate) (PS-*b*-PMMA) diblock copolymer that we used forms parallel cylindrical domains of PMMA, with a repeat spacing of 50 nm, surrounded by PS. Sample fabrication followed a two-level self-assembly process. First, spin-casting of the copolymer from solution to a thickness equivalent to one repeat spacing, and subsequent annealing, produced a thin film consisting of laterally alternating domains^{11,12} (Fig. 1). Control over the domain curvature and orientation is possible, for example by applied in-plane electric fields^{10,16}, but was not the focus of the work described here. Instead, we consider the copolymer domain pattern as a given scaffold, and focus on the second assembly step, in which a small amount of metal (of nominal thickness 0.5–12 nm) was thermally evaporated onto the copolymer film, and the metal atoms were allowed to diffuse^{17,18}. During this second step, the template morphology does not change as long as the molecular mass is sufficiently high that the order–disorder temperature is not reached¹⁰.

For a wide range of metals and small deposited amounts (film thickness typically < 0.6 nm, depending on the metal and the deposition rate), selective decoration of one of the two copolymer domains is already apparent immediately after metal deposition (Fig. 2a). We observed this behaviour for Au, Ag, In, Pb, Sn and Bi, with Au and Ag preferring the PS domain, and In, Pb, Sn and Bi the PMMA domain. In each of these cases, the preferred domain becomes the scaffold. Selectivity of nearly 100% is achieved by warming the system in an inert gas atmosphere to a temperature that is above the copolymer glass temperature, but below the point at which the template pattern disintegrates (< 300 °C for PS-*b*-PMMA) (Fig. 2b). Inside the preferred domain, this leads to a coarsening that increases both the size and the separation of the nanoparticles while the metal also diffuses into the domains, perpendicular to the film plane^{18,19}. With the exception of Ag, depositing larger amounts of metal does not fill out the domains. Rather than coalescing into continuous, elongated structures of high aspect ratio, the material ignores the template and ‘jumps the

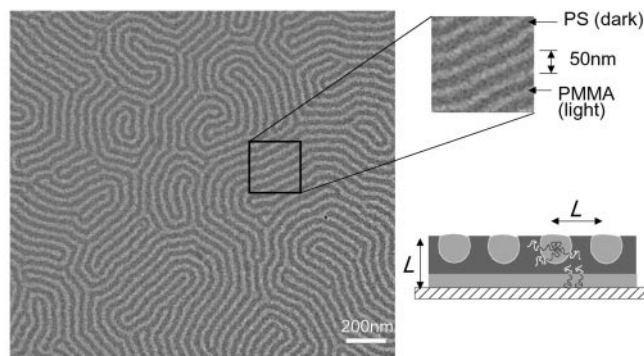


Figure 1 Transmission electron microscopy (TEM) image of ultrathin diblock copolymer film. The PS-*b*-PMMA copolymer (total molecular mass, 84,000; PS molecular mass, 60,000; polydispersity, 1.08; Soxhlet-extracted with cyclohexane to remove excess homopolymer PS in the diblock copolymer as obtained from Polysciences, Inc.) was spin-cast from solution to obtain a film thickness corresponding to the bulk repeat spacing, L . Spin-casting onto amorphous Si_3N_4 membrane substrates²⁵ allowed direct investigation by TEM without sample removal. The sketch shows a cross-section of the resulting film morphology. The striped domain pattern spontaneously forms as a consequence of phase separation between the two copolymer blocks, and provides the first level in a hierarchy of self-assembly levels. Image contrast results from electron-beam irradiation-thinning of the PMMA domains, which therefore appear lighter.

[†] Present address: Arryx Inc., 316 N. Michigan Avenue, Suite CL20, Chicago, Illinois 60601, USA.

gap' between neighbouring domains. Subsequent annealing only reduces the aspect ratio and leads to particles that are more like spheres, but does not reinstate selectivity. These results make it clear that high-density, selective metal decoration cannot result from equilibrium states of the metal–copolymer system.

One path to increased metal density is to use repeated deposition and short-time annealing of small amounts of metal. This is equivalent to generating a large number of nucleation sites for metal droplet formation inside the preferred domains, while limiting nanoparticle coalescence by keeping the coarsening far from the final, equilibrium state. This approach produces dense nanochains (Fig. 2c) of fairly narrow particle size distribution. A second possibility utilizing non-equilibrium behaviour involves tailoring the metal–polymer interactions such that, at least over some time window, the metal effectively wets one of the two copolymer domains but de-wets the other. This requires a very high mobility contrast between the two copolymer domains, allowing the metal to clear one domain and aggregate at high volume fraction in the other before surface tension drives the aggregate into its final, spherical shape. However, under a wide range of conditions (deposition of films up to 15 nm thick at rates from 0.01 to 1 nm s⁻¹), we never found Au, In, Pb, Sn or Bi to fully wet their preferred domain during any part of the self-assembly process. At best, we observed ellipsoids with aspect ratios of no more than 4–5 for Au.

The exception is silver, which exhibits behaviour very different from all the other metals that we investigated and, at coverages > 10 nm, easily forms continuous nanowires along the PS domains (Fig. 2d and e). The extremely high surface mobility of Ag on

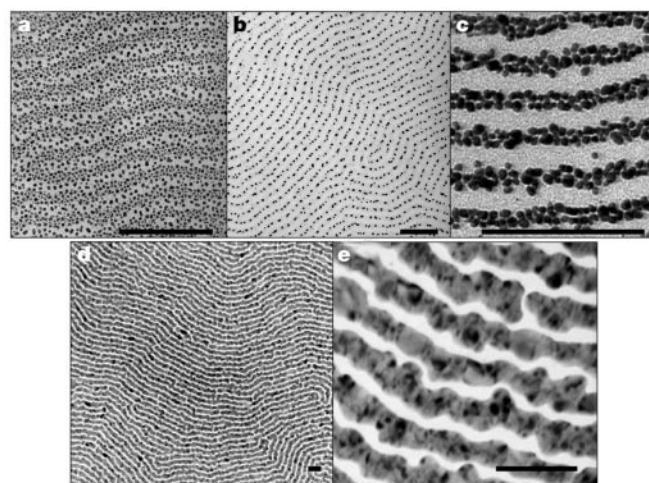


Figure 2 Metal nanochain and nanowire formation on PS-*b*-PMMA at different stages during the second self-assembly level. **a**, Aggregation of Au metal, vapour-deposited onto copolymer film (3 nm nominal metal film thickness), into nanometre-sized islands. Before annealing, the difference in wetting properties between the two copolymer domains is already apparent. **b**, Annealing at 180 °C for 1 min in an Ar atmosphere produces highly selective Au decoration of the PS domains. Inside the preferred domains (the scaffold), the metal forms chains of nanoparticles. **c**, Repeated deposition and short-time annealing (3 × 2 nm / 180 °C) leads to densely packed nanochains. With this procedure, PS domains can be loaded with metal nanoparticles to volume fractions exceeding 30% without incurring more than 10 metal bridges per square micrometre between neighbouring PS regions. Qualitatively similar results are obtained for In, Sn, Pb or Bi. **d**, Large-scale TEM micrograph, and **e**, close-up view of self-assembled Ag nanowires. Nominally 12-nm Ag was deposited onto the copolymer film without subsequent annealing. Note the extreme mobility and selectivity of Ag. Ag prefers the PS domains (see also Fig. 3b). Smaller deposited amounts of Ag (<3 nm) show the same high selectivity, but assemble into nanochains as in **b**. Individual nanocrystallites inside the wires exhibit varying grey levels owing to Bragg scattering from differently oriented lattice planes. All metal depositions were performed at base pressures of (1–2) × 10⁻⁶ torr and rates of 0.005–0.1 nm s⁻¹. Scale bars: 200 nm in **a–c**, and 100 nm in **d** and **e**.

PMMA is apparent from the fact that no annealing is required to produce essentially 100% selectivity (Fig. 2e). Changes in the rate of metal deposition did not qualitatively affect wire formation or selectivity. Note that these wires can follow the 25-nm-wide polymer scaffolds over distances of micrometres without breaks. Once formed, the wires were stable for at least 9 months at room temperature. Yet with a surface tension for Ag of around 1.2–1.4 J m⁻² (ref. 20), three orders of magnitude higher than that for the polymers²¹, and with the silver–silver interaction known to dom-

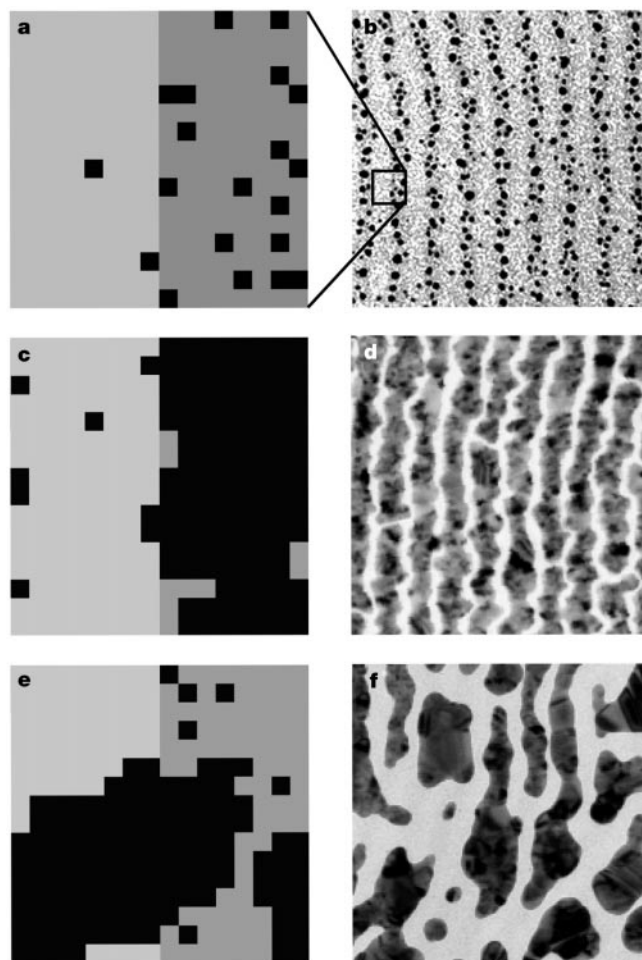


Figure 3 Comparison of Monte Carlo simulation (left column) and experimental results (right column) for Ag nanowire formation. The simulation results shown were obtained for mobility ratio $m \ll 1$ and sticking coefficient $0.8 < S < 1$. **a, b**, Early stages of metal diffusion and aggregation after 20 out of a total of 128 random walkers have been started in the simulation, or about 2 nm of Ag has been deposited in the experiment. The simulation considers diffusion and aggregation on a lattice consisting of a high- and a low-mobility side, coloured light and dark grey, respectively. The aggregation of Ag atoms into well separated nanoparticles with typical diameters around 10 nm inside the polystyrene domains, corresponding to the darker regions in the TEM micrograph, is seen in **b** and replicated in **a, c, d**. Once all 128 walkers have been randomly added to the simulation, highly domain-selective aggregation into continuous wire-like nanostructures completely fills the low-mobility side of the lattice. In the experiment, this corresponds to deposition of >12 nm Ag. **e, f**, At later stages in the simulation (**e**), the highly anisotropic, elongated wires in **c** deform. This is mediated by a strong metal–metal interaction that allows for a high mobility of walkers along the perimeter of aggregates, negating the low mobility for diffusion on the underlying polymer and favouring rounded overall shapes. Qualitatively, the same tendency toward more spherical, equilibrium shapes is seen in the experiments if the system is warmed to >80 °C (**f**). Removal of the Ag by a suitable etch shows that only the metal deforms, while the underlying template retains its original morphology as in Fig. 1 (in the TEM micrograph (**f**), the comparatively low-contrast polymer domains are not discernible).

inate over the Ag–PS interaction²², the wires are clearly non-equilibrium structures: the equilibrium state for Ag on the diblock is of spherical shape, as for the other metals.

To elucidate the kinetic mechanism responsible for wire formation, we performed two-dimensional Monte Carlo simulations. Figure 3 compares snapshots from one such simulation with experimental results. The simulation involved 16×16 lattice sites with periodic boundary conditions to represent a $50 \text{ nm} \times 50 \text{ nm}$ 'unit cell' of the sample, containing one PMMA domain (light grey)

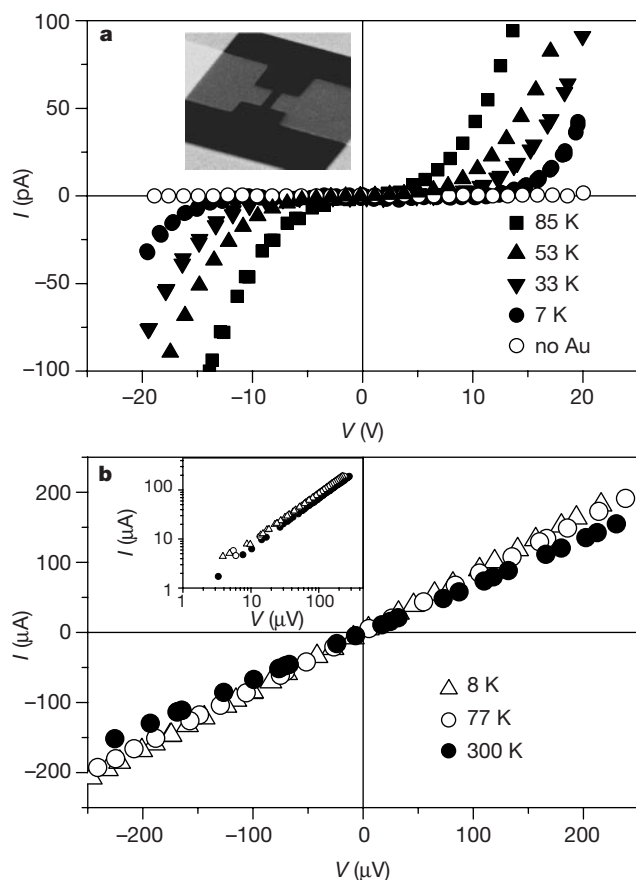


Figure 4 Electronic transport measurements on self-assembled nanochains (**a**) and nanowires (**b**). The samples were prepared on silicon nitride membrane substrates²⁵ which allowed for all self-assembly steps, as well as in-plane electrode fabrication, low-temperature transport measurements, and finally direct TEM observation to be performed on the same sample without removal from the substrate. The inset to **a** shows a scanning electron micrograph of the central region of the substrate (a membrane 100 nm thick, transparent to electrons and thus appearing dark) with 17-nm-thick Cr electrodes that were used both for copolymer domain alignment¹⁶ during the first assembly level, and later to interface the nanochain array to external circuitry. **a**, The self-assembled copolymer scaffold, not yet decorated with Au nanoparticles, exhibits current levels $< 50 \text{ fA}$ across the in-plane electrodes (gap spacing $4 \mu\text{m}$). After Au decoration (see Fig. 2c for a close-up of the sample), highly nonlinear current–voltage characteristics (I – V curves) are observed. The nonlinearity arises from large charging energies that at low bias voltages prevent electron tunnelling between nanoparticles embedded in the surrounding polystyrene matrix. The spread in nanoparticle sizes and spacings along the current paths inside each copolymer domain translates into a range of single-electron charging energies and tunnel resistances. At higher temperatures, additional paths involving smaller islands—that is, larger charging energies—become accessible and lead to an increased conductance, particularly at small bias voltages²⁶. **b**, Metallically continuous Ag nanowires, by contrast, exhibit highly linear I – V curves. The inset shows that this ohmic behaviour extends down to the smallest voltage levels measured. This sample, shown close up in Fig. 2d and e, had 2-mm-wide in-plane electrodes, resulting in an array of $\sim 10^4$ parallel nanowires of which about 10% were continuous across the electrode gap ($12 \mu\text{m}$).

and one PS domain (dark grey). Occupied sites on the lattice ('random walkers', black squares) representing Ag atoms or small Ag clusters were allowed to diffuse and aggregate over time. Approximating the way that the atoms land on the diblock template during metal deposition and the time that small clusters take to form, the simulation added individual walkers to the lattice gradually, but spatially randomly, until 50% of all available lattice sites were occupied. The ratio of the mobilities on the two halves of the lattice is $m \approx D_{\text{PS}}/D_{\text{PMMA}}$, where D_i is the metal diffusion constant for motion in copolymer domain i . The metal–metal interaction was modelled by an attractive interaction mediated by a sticking coefficient, S , and a hard-core repulsion to avoid particle overlap. The mobility of metal aggregates containing $N > 1$ neighbouring walkers was taken to vary as $1/N$. Because the metal prefers one of the copolymer domains (PS in the case of Ag or Au) over the other, even under equilibrium conditions, a small energy difference between the two halves of the lattice was included. This difference is insignificant, however, when compared to the attractive metal–metal interaction.

By a suitable choice of the two parameters m and S , we achieved a good correspondence between experiment and simulation (Fig. 3), including the early stages at very low metal coverage and the formation of continuous wires at intermediate times. Also apparent is the disintegration of these wires as the system comes closer to equilibrium. Experimentally, the self-assembled wire structures can be stabilized by quenching the sample after metal deposition (to room temperature in the case of Ag). Moderate warming to $> 80^\circ\text{C}$ changes the Ag wires into rounded shapes (Fig. 3f), which correspond well with the bulbous structures seen in the simulation at late stages (Fig. 3e). Close inspection by transmission electron microscopy (TEM) and diffraction measurements show that the Ag wires immediately after metal deposition (Fig. 3d) are highly polycrystalline, while a slight anneal induces a rapid growth of larger (tens of nanometres) single-crystalline regions before the wires disintegrate. This again demonstrates a very large surface mobility of Ag atoms, not only on PMMA but also on the Ag nanocrystallites.

Although a more quantitative comparison with experiment may require a more realistic model, the present simulation demonstrates the possibility of tailoring the copolymer + metal system to optimize wire formation. Exploration of the (m, S) parameter space shows²³ that this requires $m \rightarrow 0$ and $S \rightarrow 1$. As metal–metal bond energies E_b are large compared to the thermal energy, $k_B T$, near room temperature, the parameter $S = 1 - \exp[-E_b/(k_B T)]$ cannot deviate much from unity. However, m is affected by the choice of both copolymer and metal. From control experiments²³ with Ag on homopolymer PS and PMMA of the same molecular masses as in the copolymer we estimate²³ $0.005 < m < 0.03$. Although we found such a value only for Ag, suitable metal alloys and different copolymers should provide a straightforward means for tuning m .

To demonstrate the versatility of this hierarchical route for self-assembly of functional, laterally nanostructured systems, we now discuss the distinctly different behaviour produced by the nanochains and nanowires shown in Fig. 2. At high metal-loading fractions, the particles come close enough to provide electrically conducting paths. Figure 4 shows electronic transport measurements obtained for fully self-assembled conductors of this type. Nanochains exhibit highly nonlinear current–voltage (I – V) curves (Fig. 4a), characteristic of single-electron tunnelling between nanometre-sized metal islands in the presence of strong charging (Coulomb blockade) effects²⁴. The large overall threshold voltage for current onset results from the series combination of many such tunnel junctions along the current path. At low voltage bias, the structures are essentially insulating but switch over to conducting behaviour beyond the threshold. By contrast, the I – V curves for the nanowires are linear (Fig. 4b). Such ohmic response is indicative of continuous, metallic connections across the sample. The resistivity of a few $\mu\Omega \text{ cm}$ (using TEM on the same sample to estimate the

number and geometry of fully connected paths) is in line with the notion of grain-boundary-dominated transport in a polycrystalline metal, as is the observation of only a small decrease of the resistance with decreasing temperature. Thus, for a given overall geometry of conduction paths, controlled by the self-assembly of the copolymer scaffold, the ability to tune the metal–polymer kinetics of the second assembly level provides the flexibility to tailor the transport behaviour independently—and over the whole range from insulating or semiconducting to metallic.

This ability to separate, by hierarchical self-assembly, the formation of the organic scaffold from the processes functionalizing it with an inorganic component, in our case metal, opens up possibilities of assembling nanostructures with large anisotropy and very high local dielectric contrast, suitable for nanogratings, interconnects or sensor applications. Finally, although we have described a prototype hierarchical assembly process with two levels, additional levels are clearly possible. For example, the structures shown in Fig. 2 might form the basis for electroless plating steps or for the attachment of suitable biomolecules. □

Received 6 June; accepted 26 October 2001.

- Boal, A. K. *et al.* Self-assembly of nanoparticles into structured spherical and network aggregates. *Nature* **404**, 746–748 (2000).
- Black, C. T., Murray, C. B., Sandstrom, R. L. & Sun, S. Spin-dependent tunneling in self-assembled cobalt-nanocrystal superlattices. *Science* **290**, 1131–1134 (2000).
- Gau, H., Herminghaus, S., Lenz, P. & Lipowsky, R. Liquid morphologies on structured surfaces: from microchannels to microchips. *Science* **283**, 46–49 (1999).
- Higgins, A. M. & Jones, R. A. L. Anisotropic spinodal dewetting as a route to self-assembly of patterned surfaces. *Nature* **404**, 476–478 (2000).
- Thurn-Albrecht, T. *et al.* Ultrahigh-density nanowire arrays grown in self-assembled diblock copolymer templates. *Science* **290**, 2126–2129 (2000).
- Park, M., Harrison, C., Chaikin, P. M., Register, R. A. & Adamson, D. H. Block copolymer lithography: periodic arrays of $\sim 10^{11}$ holes in 1 square centimeter. *Science* **276**, 1401–1404 (1997).
- Spatz, J. P., Mössner, S., Hartmann, C. & Möller, M. Ordered deposition of inorganic clusters from micellar block copolymer films. *Langmuir* **16**, 407–415 (2000).
- Whitesides, G. M., Mathias, L. J. & Seto, C. T. Molecular self-assembly and nanochemistry: A chemical strategy for the synthesis of nanostructures. *Science* **254**, 1312–1319 (1991).
- Sanchez, C. & Lebeau, B. Design and properties of hybrid organic-inorganic nanocomposites for photonics. *Mater. Res. Soc. Bull.* **26**, 377–387 (2001).
- Fasolka, M. J. & Mayes, A. M. Block copolymer thin films: physics and applications. *Annu. Rev. Mater. Res.* **31**, 323–355 (2001).
- Morkved, T. L. & Jaeger, H. M. Thickness-induced morphology changes in lamellar diblock copolymer ultrathin films. *Europhys. Lett.* **40**, 643–648 (1997).
- Tang, W. H. & Witten, T. A. Quenched degrees of freedom in symmetric diblock copolymer thin films. *Macromolecules* **31**, 3130–3135 (1998).
- Harrison, C. *et al.* Mechanisms of ordering in striped patterns. *Science* **290**, 1558–1560 (2001).
- Balazs, A. C. Interactions of nanoscopic particles with phase-separating polymeric mixtures. *Curr. Opin. Colloid Interface Sci.* **4**, 443–448 (2000).
- Morkved, T. L., Wiltzius, P., Jaeger, H. M., Grier, D. G. & Witten, T. A. Mesoscopic self-assembly of gold islands on diblock-copolymer films. *Appl. Phys. Lett.* **64**, 422–424 (1994).
- Morkved, T. L. *et al.* Local control of microdomain orientation in diblock copolymer thin films with electric fields. *Science* **273**, 931–933 (1996).
- Cole, D. H., Shull, K. R., Rehn, L. & Baldo, P. Metal-polymer interactions in a polymer/metal nanocomposite. *Phys. Rev. Lett.* **78**, 5006–5009 (1997).
- Lin, B. *et al.* X-ray studies of polymer/gold nanocomposites. *J. Appl. Phys.* **85**, 3180–3184 (1999).
- Guico, R. S., Richter, A. G., Wang, J. & Shull, K. R. X-ray standing wave measurements of gold nanoparticles in polymeric thin films. *Polymeric Materials: Science & Engineering (Abst.)* **85**, (American Chemical Society, Washington, in the press).
- Piuz, F. & Borel, J. P. Thermodynamical size effect in small particles of silver. *Phys. Status Solidi A* **14**, 129–133 (1972).
- Wu, S. *Polymer Interfaces and Adhesion* (Marcel Dekker, New York, 1982).
- Gerenser, L. J. & Goppert-Berarducci, K. E. in *Metalized Plastics 3: Fundamental and Applied Aspects* (ed. Mittal, K. L.) 163–178 (Plenum, New York, 1992).
- Lopes, W. Non-equilibrium self-assembly of metals on diblock copolymer templates. *Phys. Rev. E* (submitted).
- Ford, E. M. & Ahmed, H. Control of Coulomb blockade characteristics with dot size and density in planar metallic multiple tunnel junctions. *Appl. Phys. Lett.* **75**, 421–423 (1999).
- Morkved, T. L., Lopes, W. A., Hahm, J., Sibener, S. J. & Jaeger, H. M. Silicon nitride membrane substrates for the investigation of local structure in polymer thin films. *Polymer* **39**, 3871–3875 (1998).
- Cordan, A. S. *et al.* Temperature behavior of multiple tunnel junction devices based on disordered dot arrays. *J. Appl. Phys.* **87**, 345–352 (2000).

Acknowledgements

We thank T. Witten, S. Coppersmith, T. Morkved, M. Möller and K. Shull for discussions, and R. Parthasarathy for help with some of the *I*–*V* measurements. This work was supported by the MRSEC programme of the NSF and by the W. M. Keck Foundation.

Correspondence should be addressed to H.M.J. (e-mail: h-jaeger@uchicago.edu).

Himalayan tectonics explained by extrusion of a low-viscosity crustal channel coupled to focused surface denudation

C. Beaumont*, R. A. Jamieson†, M. H. Nguyen*† & B. Lee*

* Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada, B3H 4J1

† Department of Earth Sciences, Dalhousie University, Halifax, Nova Scotia, Canada, B3H 3J5

Recent interpretations of Himalayan–Tibetan tectonics have proposed that channel flow in the middle to lower crust can explain outward growth of the Tibetan plateau^{1–3}, and that ductile extrusion of high-grade metamorphic rocks between coeval normal- and thrust-sense shear zones can explain exhumation of the Greater Himalayan sequence^{4–7}. Here we use coupled thermal–mechanical numerical models to show that these two processes—channel flow and ductile extrusion—may be dynamically linked through the effects of surface denudation focused at the edge of a plateau that is underlain by low-viscosity material. Our models provide an internally self-consistent explanation for many observed features of the Himalayan–Tibetan system^{8–10}.

Figure 1 shows a general cross-section, highlighting features of the orogen (numbers 1–9) requiring explanation in any quantitative model¹⁰. These include contemporaneous north–south shortening on the Main Central thrust (MCT) system (1) and extension along the South Tibetan detachment (STD) system (2), the creation of gneiss domes (3) in the southern Tibetan plateau, the style and timing of metamorphism in the Greater Himalayan (GHS, 5) and Lesser Himalayan (LHS, 6) sequences, and the juxtaposition of contrasting protoliths (8) across the MCT zone. We show below that these disparate observations can all be interpreted to result from coupling between channel flow originating in partially molten crust beneath Tibet (4) and rapid denudation of the south flank of the Himalaya (7).

We first present the behaviour of a reference model (model 1; Fig. 2), and demonstrate a similar tectonic style with boundary

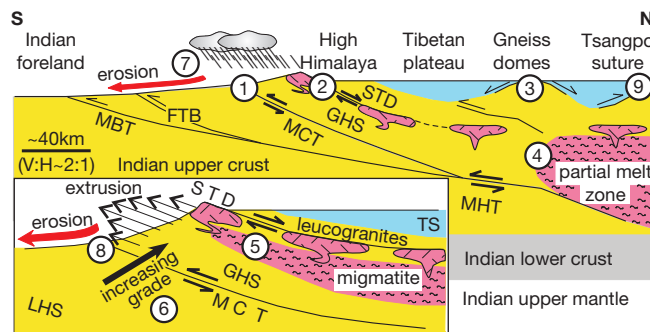


Figure 1 General tectonic features of the Himalaya and southern Tibet^{5,8–10,15,16}. Numbers 1–9 correspond to important features requiring explanation¹⁰; see Figs 2–4 for corresponding features of models. Inset shows southern flank of Tibetan plateau (not to scale) with geological features discussed in text. LHS, Lesser Himalayan sequence; GHS, Greater Himalayan sequence; TS, Tethyan sequence; MCT, Main Central thrust; STD, South Tibetan detachment; MHT, Main Himalayan thrust; MBT, Main Boundary thrust; FTB, fold-thrust belt (Siwaliks). Colours (all figures): blue, weak upper crust; yellow, medium-strength middle crust; grey, strong lower crust; pink, melt-weakened middle crust (includes migmatite and plutons).

conditions appropriate for the Himalayan–Tibetan system (model 2; Fig. 2). The sensitivity of the model results to surface denudation rate and upper-crust rheology is investigated (Fig. 3), and results compatible with observations from a generalized Himalayan cross-section are presented (model 3, Fig. 4). The models build on our earlier mechanical¹¹ and thermal–mechanical^{12,13} results. In the context of the Himalayan–Tibetan system (Fig. 1), ‘pro’-lithosphere, corresponding to India, converges on stationary ‘retro’-

lithosphere, corresponding to Asia, at velocity V_p . Pro-mantle lithosphere is detached and subducted at point S (Fig. 2; details in legend and Supplementary Information); the model ‘suture’ is the initial boundary between pro- and retro-crust. Surface denudation rate depends on local surface slope, time, and a spatial climate function^{12,13} (Fig. 2; details in legend) that yields high denudation rates on the steep pro-ward flank of the plateau, corresponding to the Himalaya, and low-to-negligible rates elsewhere. The most

Model 1: Development of crustal channel/extrusion

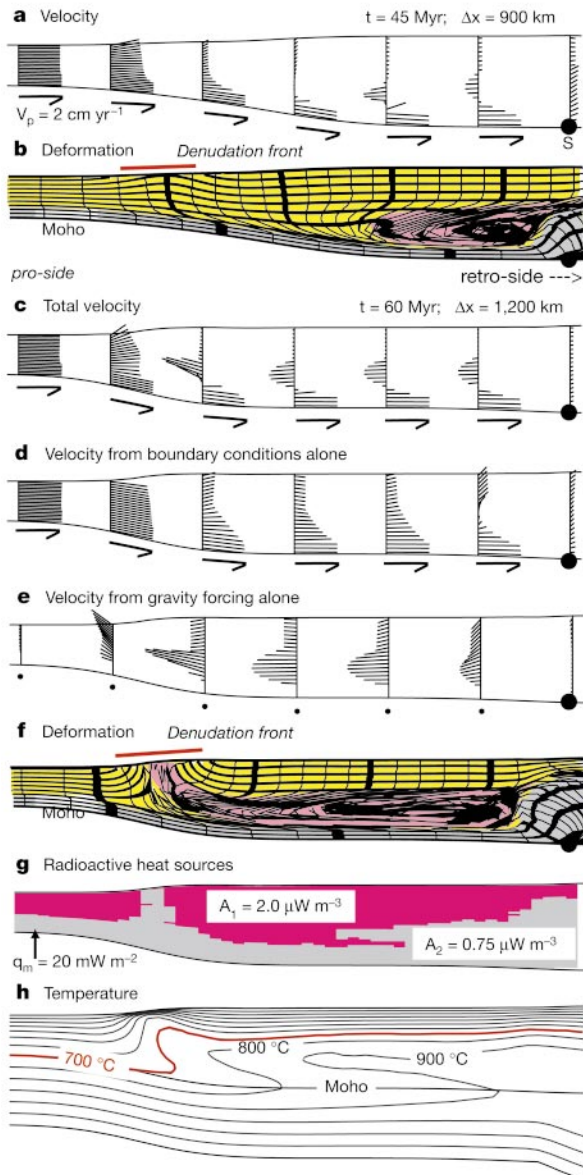


Figure 2 Development of surface-coupled channel flow (a–h) and effect of changing convergence velocity and subduction conditions (i–j). Pro-side is to left in all panels; retro-side lies to right of S and is generally not shown; see also Supplementary Information. Colours as for Fig. 1 unless otherwise noted. In all figures, times are in Myr after start of the model unless otherwise indicated. Model 1 ($V_p = 2 \text{ cm yr}^{-1}$; $V_s = 0$; no crust subducted): **a**, Crustal velocity field at 45 Myr, showing reverse flow in mid-crust, basal velocity (arrows) and detachment point (S); **b**, deformed passive marker grid (45 Myr), showing ‘tunnelling’ of channel before coupling to surface denudation; **c**, crustal velocity field at 60 Myr, showing reverse channel flow coupled to surface denudation; **d**, retro-ward velocity resulting from basal boundary conditions alone; **e**, pro-ward velocity resulting from gravitational forcing; **f**, deformed marker grid (60 Myr) showing extrusion of channel at denudation front; **g**, corresponding distribution of radioactive material in crust (magenta, A_1 , high heat production; grey, A_2 , lower heat production); **h**, crustal isotherms, showing lateral advection of hot material in channel; 700 °C isotherm (red) marks onset of melt weakening. Model 2 shows effect of lower crustal subduction and partitioning of basal velocity, $V_p = 5 \text{ cm yr}^{-1}$, into subduction and subduction zone advance (V_s at 2.5 cm yr^{-1} each (V_s is the rate the subduction point S moves in the retro-direction)): **i**, crustal velocity field at 51 Myr; **j**, deformed marker grid at same time, showing channel extrusion as in **f** and position of suture (9). Heat balance equation: $\rho C_p (\partial T / \partial t + v \cdot \nabla T) = K \nabla^2 T + A$; thermal parameters (same for all models, see also **g**): asthenospheric temperature $T_a = 1,350 \text{ °C}$; specific heat at constant pressure $C_p = 750 \text{ m}^2 \text{ K}^{-1} \text{ s}^{-2}$; $\rho_{\text{crust}} = 2,700 \text{ kg m}^{-3}$; $\rho_{\text{mantle}} = 3,300 \text{ kg m}^{-3}$; thermal conductivity $K = 2.0 \text{ W m}^{-1} \text{ K}^{-1}$; thermal diffusivity $\kappa = 1.0 \times 10^{-6} \text{ m}^2 \text{ s}^{-1}$; A is the volumetric radioactive heat production. Flexural rigidity, 10^{22} N m . Ductile flow law: $\eta = B^* (\dot{\epsilon})^{1/n-1} \exp(Q^*/nRT)$, where η is effective viscosity, B^* is the pre-exponential coefficient, n is the power-law coefficient, $\dot{\epsilon}$ is the second invariant deviatoric strain rate, and Q^* is the activation energy. Coulomb (frictional) flow law: effective internal angle of friction, $\phi = 15^\circ$. Upper-crust rheology (0–25 km): ‘wet’ Black Hills quartzite¹⁴, $n = 4.0$, $Q^* = 223 \text{ kJ mol}^{-1}$, $B^* = 2.92 \times 10^6 \text{ Pa s}^{1/4}$; ‘melt weakening’: effective viscosity decreases from flow law value at 700 °C to $\eta = 10^{19} \text{ Pa s}$ at $T \geq 750 \text{ °C}$; lower-crust rheology (25–35 km): ‘dry’ Maryland diabase³⁰ with $B^*/10$, $n = 4.7$, $Q^* = 485 \text{ kJ mol}^{-1}$, $B^* = 1.91 \times 10^5 \text{ Pa s}^{1/4.7}$. Denudation rate: 0 for $t < 37.5 \text{ Myr}$; for $t > 37.5 \text{ Myr} = 0.107 \times \text{slope} \times g(x) \text{ m yr}^{-1}$, where $g(x) = 0$ at $x > 500 \text{ km}$, $g(x) = 1$ at $x < 450 \text{ km}$, linear decrease in $g(x)$ from 1 to 0 between 450–500 km. This represents humid to arid transition between foreland and plateau, and gives denudation rates of about 1 cm yr^{-1} for slopes of 1:10 (averaged on 10 km horizontal scale).

important model property is an extra increment of weakening in the upper crust^{12,13} such that effective viscosity is reduced linearly from the power-law flow value¹⁴ at 700 °C to 10^{19} Pa s at $T \geq 750$ °C, a factor of about 10. This is probably a conservative estimate for weakening by a small amount of *in situ* partial melt, here referred to as 'melt weakening'.

Figure 2 illustrates interactions among temperature, rheology, channel flow, and surface denudation. Model 1 (basic subduction) has $V_p = 2$ cm yr⁻¹, whereas model 2 (advancing subduction) has $V_p = 5$ cm yr⁻¹, $V_s = 2.5$ cm yr⁻¹ (Fig. 2), and includes subduction of lower pro-crust. Initial conditions and early evolution are described elsewhere¹³ (see Supplementary Information). Note that despite the different boundary conditions, both models have similar channel-denudation interactions at the denudation front (Fig. 2f and j). For large convergence ($\Delta x > 1,000$ km), the 70–75-km-thick crust has a well developed plateau, and much of the middle to lower crust is above 700 °C. Reverse channel flow (Fig. 2c and i) develops where hot upper crust is melt weakened, and is subject to a pressure gradient developed between the plateau and the foreland^{1–3,12,13}.

Initially, the channel flow propagates ('tunnels') through the converging pro-crust (Fig. 2a and b) at a rate controlled by the temperature increase and associated viscosity decrease at the channel tip^{12,13}. By $\Delta x = 1,200$ km, efficient erosion of the pro-flank of the plateau leads to coupling between channel flow and surface denudation (Fig. 2f). Channel velocity increases, and the channel is exhumed at the erosion front in a narrow zone bounded by upper normal-sense and lower thrust-sense shear zones (Fig. 2f). Isotherms are advected pro-ward by channel flow, and condensed near the surface by cooling accompanying denudation (Fig. 2g and h). Decomposing the total velocity field (Fig. 2c) into tectonic and

gravitational components¹² (Fig. 2d and e) shows that gravity forces drive the channel flow.

The two models have similar thermal–tectonic styles in the region that corresponds to the Himalaya and southern Tibetan plateau, demonstrating relative insensitivity to the choice of boundary conditions. Model 2 ($V_p = 5$ cm yr⁻¹, $V_s = 2.5$ cm yr⁻¹) is more compatible with the average 5 cm yr⁻¹ post-collision convergence rate of India relative to Asia^{10,15}, and at 51 Myr also has properties consistent with observed crustal thickness, plateau width, suture position^{15,16} (Fig. 2i and j), extent of underthrust pro-mantle (Indian) lithosphere as inferred seismically¹⁶, and temperature distribution (similar to Fig. 2h). However, the broadly symmetrical geometry of both models at the denudation front (Fig. 2f and j) contrasts with typical asymmetrical Himalayan structures. Factors affecting the range of model styles, including symmetry, are addressed below.

The model results are sensitive to variations in surface denudation rate and upper-crust rheology, leading to a variety of predicted flow modes (Fig. 3; details in legend). Changes in denudation rate affect the degree of coupling between the surface and the channel, and the presence of a weak upper-crust layer, analogous to a sedimentary sequence overlying crystalline crust, affects the mechanical behaviour of the crust above the channel. Although melt weakening to 10^{19} Pa s is required to initiate and sustain channel flow, further viscosity reduction to 10^{18} Pa s does not strongly influence the model results¹². Some of the predicted flow modes may be comparable to different regions or evolutionary stages of the Himalayan–Tibetan system.

In the absence of denudation, the channel flow tunnels through converging crust^{12,13} (Fig. 3a), as has been proposed to explain eastward propagation of the Tibetan plateau^{1–3}. Denudation

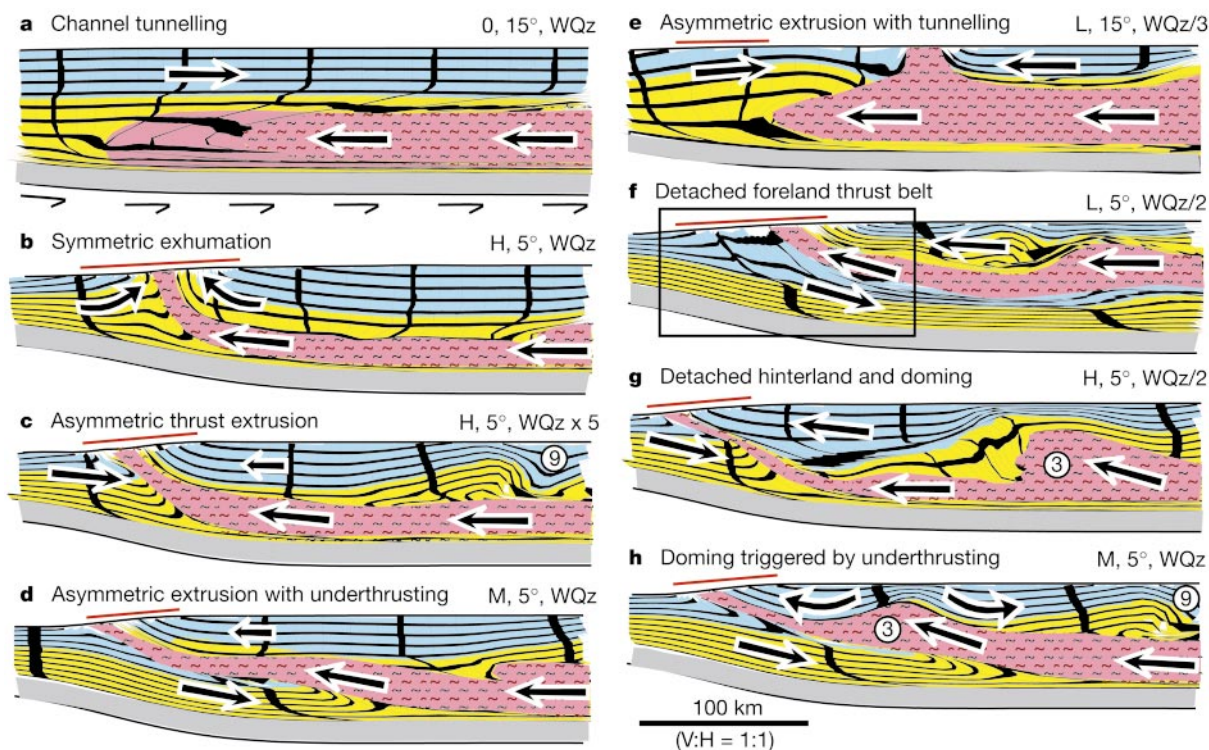


Figure 3 Model channel flow modes and exhumation/extrusion styles. Properties as for model 2 except for upper-crust rheology (blue; $\phi = 5^\circ$ or 15° ; WQz, wet quartzite flow law¹⁴, effective viscosity scaled as shown) and erosion rate, which ranges from 0 to low (L, < 0.4 cm yr⁻¹), moderate (M, 0.4 – 1.4 cm yr⁻¹), or high (H, > 1.4 cm yr⁻¹) where surface slope is maximum for the time step shown. **a**, Channel tunnelling, no denudation^{12,13}; **b**, symmetric exhumation (models 1, 2); **c**, asymmetric thrust extrusion; **d**, asymmetric extrusion with underthrusting; **e**, asymmetric extrusion with tunnelling; abandoned

extrusion zone translated into plateau; **f**, detached foreland fold-thrust belt; box outlines region similar to that shown in Fig. 4; **g**, detached hinterland followed by doming and extension; **h**, dome triggered by underthrusting. Colours as for Fig. 1; red line above model surface indicates denudation front; pink indicates melt-weakened channel flow zone (strongly deformed grid removed for clarity); arrows show flow direction; (3) indicates structures resembling north Himalayan gneiss domes¹⁹, (9) indicates suture.

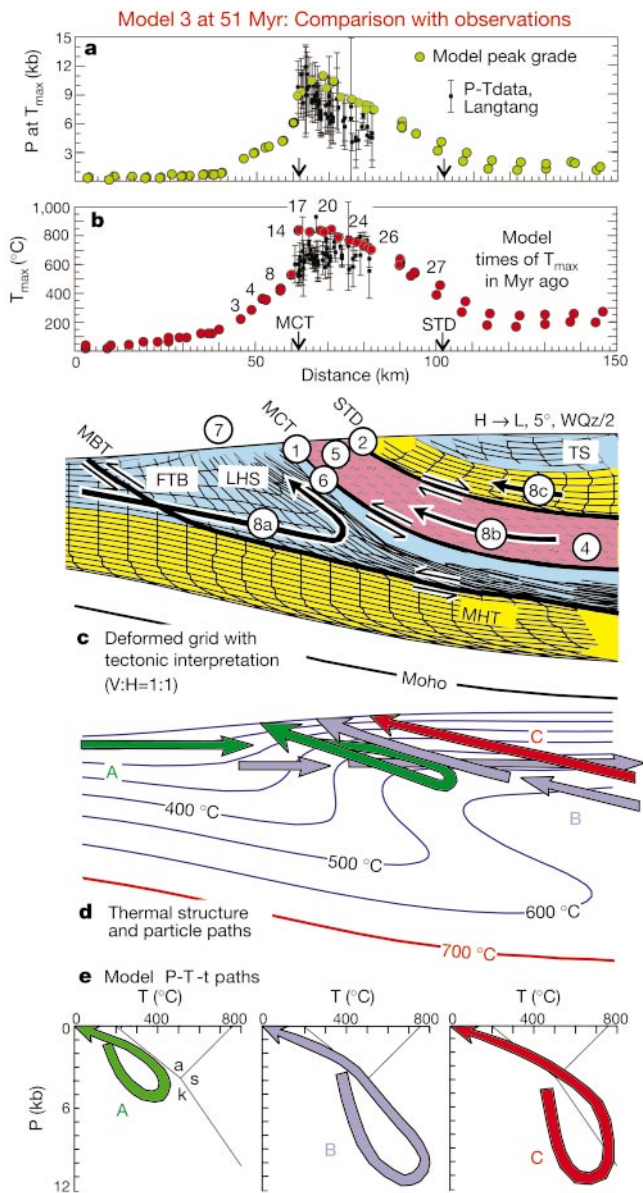


Figure 4 Comparison of model 3 (51 Myr after start of model) with observations from the Himalayan-Tibetan system. Large-scale features (for example, positions of suture and S-point) are similar to model 2 (Fig. 2j). **a**, Pressure P at $T_{\max}$ profile (green) compared with P data from the Langtang region, central Nepal^{25,26,27}, where $T_{\max}$ is the maximum temperature reached by the points shown during the evolution of the model; arrows show observed positions of the MCT and STD; **b**, $T_{\max}$ profile (red) compared with T data from the same area; numbers are model times of $T_{\max}$ in Myr ago; **c**, deformed marker grid and tectonic interpretation (see text); numbers, colours, and abbreviations as in Fig. 1; rheology and erosion rates as in Fig. 3; 8a, 8b, 8c represent contrasting rock units following trajectories shown by black arrows; **d**, isotherms at 51 Myr and particle paths (green, purple, red arrows) for points originating in different parts of the model that are exhumed together at about 51 Myr; note that orogen and isotherms advance toward foreland with time; **e**, model P - T - t paths for points shown in **d**; reference lines show stability fields of andalusite (a), kyanite (k), and sillimanite (s). Panels **a**-**d** aligned at MCT; horizontal distance between MCT and STD is sensitive to local structure and topography, but model channel thickness (10–15 km) is similar to inferred structural thickness of GHS^{7,28}.

focused on the plateau flank leads to symmetrical exhumation when the upper crust does not detach above the channel (models 1, 2; Fig. 3b), possibly analogous to the Nanga Parbat region¹⁷. Detachment of weak upper crust and its outward flow with the channel creates asymmetric thrust extrusion, as observed in the Himalaya^{7,10} (Fig. 3c). Underthrusting of cold crust beneath the exhumation front and into the channel (Fig. 3d) creates a master thrust zone between underthrust crust and the channel, analogous to the inferred Main Himalayan thrust^{9,15,16} (MHT, Figs 1, 4c). A significant reduction in denudation rate can lead to abandonment of the extrusion zone and re-establishment of channel tunnelling (Fig. 3e). For sufficiently weak upper crust and low erosion rates, the foreland detaches to form a fold-thrust belt, analogous to the Siwaliks¹⁸, in front of the exhuming channel (Fig. 3f). Detachment of weak upper crust on the hinterland side of the erosion front leads to doming of the channel beneath the extending and thinning upper crust (Fig. 3g); doming may also be triggered when underthrust crust (Fig. 3d) uplifts and squeezes the channel, causing extension in destabilized upper crust (Fig. 3h). Domes created in this way lie between the suture (9) and the denudation front, and may be equivalent to gneiss domes (3), including the Kangmar dome¹⁹, associated with the north Himalayan antiform¹⁰.

Figure 4 compares the results of a model with weak upper crust and variable denudation rate (model 3) to observations from the Himalaya of central Nepal (corresponding numbers, Figs 1 and 4). Early stages of model 3 (not shown) are similar to models 1 and 2. Channel flow is initiated by the development of low-viscosity material in the hot crust beneath the plateau (4), as has been inferred for Tibet⁹. Coeval thrust- and normal-sense distributed shear across the lower and upper parts of the channel (Fig. 4c) are interpreted to correspond to top-to-the-south motion on the MCT (1) and top-to-the-north motion along the STD (2), respectively. Coupling between focused surface denudation and channel flow leads to ductile extrusion of hot channel material, represented in nature by the high-grade migmatitic gneisses of the GHS (5) lying between the MCT and STD^{4–7}. Focused surface denudation (7) of the GHS on the steep windward flank of the Tibetan plateau is required to exhume the channel, consistent with isotopic and detrital mineral data indicating that the GHS or equivalent rocks have been actively eroded since the early Miocene^{18,20,21}.

Processes associated with channel extrusion can explain many features of the MCT zone, including the 'inverted' metamorphic profile¹⁰ (6), contrasting metamorphic ages across the MCT^{10,22}, and contrasts in age and provenance of GHS and LHS protoliths^{23,24} (8). Model peak grade profiles (Fig. 4a and b) agree well with pressure-temperature (P - T) data from the MCT zone and GHS in central Nepal^{25–27}. In model 3, denudation rate progressively decreases during the last 15 Myr, as inferred for the Himalaya^{20,28}; this produces an age pattern (Fig. 4b) that is compatible with young ages (<10 Myr ago) reported from the LHS²² and generally older ages (>15 Myr ago) from the GHS^{28,29}. The peak grade profiles and age distributions result from the juxtaposition of contrasting rock packages (8a, b, c, Fig. 4c) across the lower boundary of the channel and from distributed ductile shear within the channel and its footwall. Selected particle trajectories and P - T - t paths (Fig. 4d and e) illustrate how points originating in different parts of the system, and following different trajectories through the evolving thermal structure of the orogen, can be exhumed together at the orogenic front. These points correspond to (A) LHS rocks buried in the channel footwall and exhumed adjacent to the channel, (B) lower GHS rocks transported into the orogen and later extruded by reverse channel flow, and (C) upper GHS rocks buried near the onset of orogenesis and extruded in the upper part of the channel. In the model, crustal segments initially separated by more than 600 km (8a, 8b; Fig. 4c) have been juxtaposed across the MCT. By implication, the MCT is both a structural and protolith boundary^{4,7,23,24}.

These models represent, to our knowledge, the first quantitative

demonstration that flow in a low-viscosity crustal channel that is coupled to surface denudation provides an internally consistent explanation not only for ductile extrusion of the GHS but for many other salient features of the Himalayan–Tibetan system. The critical factors are the presence of low-viscosity material in the middle to lower crust, a variation in crustal thickness between plateau and foreland, and surface denudation that is focused on the plateau flank. The range of model styles, and by implication the tectonics of natural orogens, is sensitive to variations in denudation rate and upper-crust strength. □

Received 20 April; accepted 2 November 2001.

- Royden, L. H. Coupling and decoupling of crust and mantle in convergent orogens: implications for strain partitioning in the crust. *J. Geophys. Res.* **101**, 17679–17705 (1996).
- Clark, M. K. & Royden, L. H. Topographic ooze: Building the eastern margin of Tibet by lower crustal flow. *Geology* **28**, 703–706 (2000).
- Shen, F., Royden, L. H. & Burchfiel, B. C. Large-scale crustal deformation of the Tibetan Plateau. *J. Geophys. Res.* **106**, 6793–6816 (2001).
- Grujic, D. *et al.* Ductile extrusion of the Higher Himalayan Crystalline in Bhutan: evidence from quartz microfabrics. *Tectonophysics* **260**, 21–43 (1996).
- Wu, C. *et al.* Yadong cross structure and the South Tibetan Detachment in the east central Himalaya (89°–90°E). *Tectonics* **17**, 28–45 (1998).
- Vannay, J.-C. & Grasemann, B. Himalayan inverted metamorphism and syn-convergence extension as a consequence of a general shear extrusion. *Geol. Mag.* **138**, 253–276 (2001).
- Grujic, D., Hollister, L. S. & Parrish, R. R. Himalayan metamorphic sequence as an orogenic channel: insight from Bhutan. *Earth Planet. Sci. Lett.* (in the press).
- Burchfiel, B. C. *et al.* The South Tibetan detachment system, Himalayan orogen: Extension contemporaneous with and parallel to shortening in a collisional mountain belt. *Spec. Pap. Geol. Soc. Am.* **269**, (1992).
- Nelson, K. D. *et al.* Partially molten middle crust beneath southern Tibet: a synthesis of Project INDEPTH results. *Science* **274**, 1684–1688 (1996).
- Hodges, K. V. Tectonics of the Himalaya and southern Tibet from two perspectives. *Geol. Soc. Am. Bull.* **112**, 324–350 (2000).
- Willett, S. D., Beaumont, C. & Fullsack, P. Mechanical model for the tectonics of doubly vergent compressional orogens. *Geology* **21**, 371–374 (1993).
- Beaumont, C., Jamieson, R. A., Nguyen, M. H. & Lee, B. In *Slave - Northern Cordillera Lithospheric Evolution (SNORCLE) and Cordilleran Tectonics Workshop* (eds Cook, F. & Erdmer, P.) 112–170 (Lithoprobe Report 79, Lithoprobe Secretariat, University of British Columbia, Vancouver, 2001).
- Jamieson, R. A., Beaumont, C., Nguyen, M. H. & Lee, B. Interaction of metamorphism, deformation, and exhumation in large convergent orogens. *J. Metamorph. Geol.* **20**, 1–16 (2002).
- Gleason, G. C. & Tullis, J. A flow law for dislocation creep of quartz aggregates determined with the molten salt cell. *Tectonophysics* **247**, 1–23 (1995).
- Hauck, M. L., Nelson, K. D., Brown, L. D., Zhao, W. & Ross, A. R. Crustal structure of the Himalayan orogen at ~90° east longitude from Project INDEPTH deep reflection profiles. *Tectonics* **17**, 481–500 (1998).
- Owens, T. J. & Zandt, G. Implications of crustal property variations for models of Tibetan plateau evolution. *Nature* **387**, 37–43 (1997).
- Zeitler, P. K. *et al.* Crustal reworking at Nanga Parbat, Pakistan: Metamorphic consequences of thermal-mechanical coupling facilitated by erosion. *Tectonics* **20**, 712–718 (2001).
- DeCelles, P. G. *et al.* Stratigraphy, structure, and tectonic evolution of the Himalayan fold-thrust belt in western Nepal. *Tectonics* **20**, 487–509 (2001).
- Lee, J. *et al.* Evolution of the Kangmar Dome, southern Tibet: Structural, petrologic, and thermochronologic constraints. *Tectonics* **19**, 872–895 (2000).
- White, N. M. *et al.* Constraints on the structural evolution, exhumation, and erosion of the High Himalayan Slab, NW India, from foreland basin deposits. *Earth Planet. Sci. Lett.* (in the press).
- France-Lanord, C., Derry, L. & Michard, A. In *Himalayan Tectonics* (eds Treloar, P. J. & Searle, M. P.) 603–621 (Spec. Publ. 74, Geological Society, London, 1993).
- Harrison, T. M. *et al.* A Late Miocene-Pliocene origin for the central Himalayan inverted metamorphism. *Earth Planet. Sci. Lett.* **146**, E1–E7 (1997).
- Parrish, R. R. & Hodges, K. V. Isotopic constraints on the age and provenance of the Lesser and Greater Himalayan sequences, Nepalese Himalaya. *Geol. Soc. Am. Bull.* **108**, 904–911 (1996).
- DeCelles, P. G., Gehrels, G. E., Quade, J., LaReau, B. & Spurlin, M. Tectonic implications of U-Pb zircon ages of the Himalayan orogenic belt in Nepal. *Science* **288**, 497–499 (2000).
- Inger, S. & Harris, N. B. W. Tectonothermal evolution of the High Himalayan Crystalline Sequence, Langtang Valley, northern Nepal. *J. Metamorph. Geol.* **10**, 439–452 (1992).
- Macfarlane, A. M. An evaluation of the inverted metamorphic gradient at Langtang National Park, central Nepal Himalaya. *J. Metamorph. Geol.* **13**, 595–612 (1995).
- Fraser, G., Worley, B. & Sandiford, M. High-precision geothermobarometry across the High Himalayan metamorphic sequence, Langtang Valley, Nepal. *J. Metamorph. Geol.* **18**, 665–682 (2000).
- Searle, M. P. *et al.* Shisha Pangma leucogranite, South Tibetan Himalaya: Field relations, geochemistry, age, origin, and emplacement. *J. Geol.* **105**, 307–326 (1997).
- Hodges, K. V., Parrish, R. R. & Searle, M. P. Tectonic evolution of the central Annapurna range, Nepalese Himalaya. *Tectonics* **15**, 1264–1291 (1996).
- Mackwell, S. J., Zimmerman, M. E. & Kohlstedt, D. L. High-temperature deformation of dry diabase with application to tectonics on Venus. *J. Geophys. Res.* **103**, 975–984 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

This research was funded by Lithoprobe Supporting Geoscience and NSERC Research grants to C.B. and R.A.J., and the Inco Fellowship of the Canadian Institute for Advanced

Research to C.B. All the models were run using the finite element thermal-mechanical program developed by P. Fullsack. The work benefited from discussions with J. Braun, L. Brown, L. Derry, P. Fullsack, D. Grujic, D. Nelson, S. Medvedev, O. Vanderhaeghe and K. Whipple. Comments by L. Royden substantially improved the manuscript.

Correspondence and requests for materials should be addressed to C.B. (e-mail: Chris.Beaumont@Dal.Ca).

Effect of acoustic clutter on prey detection by bats

Raphaël Arlettaz^{†‡}, Gareth Jones[‡] & Paul A. Racey[§]

^{*} Division of Conservation Biology, Zoological Institute, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland

[†] Institute of Ecology, University of Lausanne, CH-1015 Lausanne, Switzerland

[‡] School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK

[§] Department of Zoology, University of Aberdeen, Tillydrone Avenue, Aberdeen AB24 2TZ, UK

Bats that capture animal prey from substrates often emit characteristic echolocation calls that are short-duration, frequency-modulated (FM) and broadband¹. Such calls seem to be suited to locating prey in uncluttered habitats, including flying prey, but may be less effective for finding prey among cluttered backgrounds because echoes reflecting from the substrate mask the acoustic signature of prey^{2–4}. Perhaps these call designs serve primarily for spatial orientation^{5–7}. Furthermore, it has been unclear whether the acoustic image conveyed by FM echoes enables fine texture discrimination^{3,8,9}, or whether gleaning bats that forage in echo-cluttering environments must locate prey by using other cues, such as prey-generated sounds^{5–7,10–13}. Here we show that two species of insectivorous gleaning bats perform badly when compelled to detect silent and immobile prey in clutter, but are very efficient at capturing noisy prey items among highly cluttered backgrounds, and both dead or live prey in uncluttered habitats. These findings suggest that the short, broadband FM echolocation calls associated with gleaning bats are not adapted to detecting prey in clutter.

Two major and non exclusive^{14–17} foraging tactics can be distinguished among insectivorous bats: aerial hawking (that is, the capture of airborne prey) and substrate gleaning. About one third of all microchiropteran bat species capture prey from substrates¹⁵. Unlike aerial-hawking bats that include longer-duration, and almost constant-frequency components in their echolocation calls, gleaning species emit calls that are often of low intensity and which often sweep from high to low frequencies (frequency-modulated (FM) calls) in a few milliseconds^{1,4}. A major outstanding problem in echolocation biology is the extent to which these calls are used for distinguishing prey items from substrates, particularly when the substrate is complex and generates a lot of echo clutter³ (that is, echoes from objects other than the target of interest). Some bat species emit calls at a high repetition rate ('feeding buzzes') to localize aerial prey, but switch off echolocation immediately before taking prey from surfaces: the bats may then listen instead for prey-generated sounds^{7,11}. Most experiments on gleaning bats have investigated prey detection on simple surfaces, where background echoes may not mask prey echoes. In such situations, bats may still use echolocation to detect prey¹⁸. Bats may also use echolocation to detect prey positioned close to flat surfaces¹⁹, and may even detect flying insects in grass by monitoring the insect's movement over successive echoes²⁰.

This study attempts to challenge the 'spectral image' paradigm⁹ using a behavioural approach. We predicted that bats would use echolocation to locate aerial prey, and for trawling prey from smooth surfaces (mimicking water) with their feet. If FM echolocation is poorly adapted for prey detection in clutter, however, we predicted that bats would not use it for locating prey hidden in leaf litter because echoes from prey are masked by echoes from clutter. As a model species, we chose the sibling mouse-eared bats *Myotis myotis* and *Myotis blythii* because their ecology is especially well documented^{15,21,22} and both emit FM calls with similar designs in every habitat configuration: while searching for prey, they emit calls that sweep from about 120–30 kHz in about 2 ms²³.

In the laboratory, wild-captured bats were provided with four artificial microhabitats, each mimicking foraging conditions faced by bats in nature^{15,22}: (1) a complex, highly cluttered background provided by leaf litter (as in a deciduous forest); (2) a simpler, less cluttered background provided by an artificial lawn (as in a freshly mown meadow); (3) an acoustic mirror provided by a horizontal perspex plate (as with a still water surface); (4) no immediate background provided by prey items in the air (airborne prey). The ability of the bats to detect and capture live and dead prey items, respectively, was measured under these circumstances. The echolocation calls emitted by the bats were also recorded in each situation. Examples of video and sound sequences of foraging mouse-eared bats can be downloaded from our website (<http://www.bio.bris.ac.uk/research/bats/myotis.htm>).

A first experiment on detection cues attempted to establish how

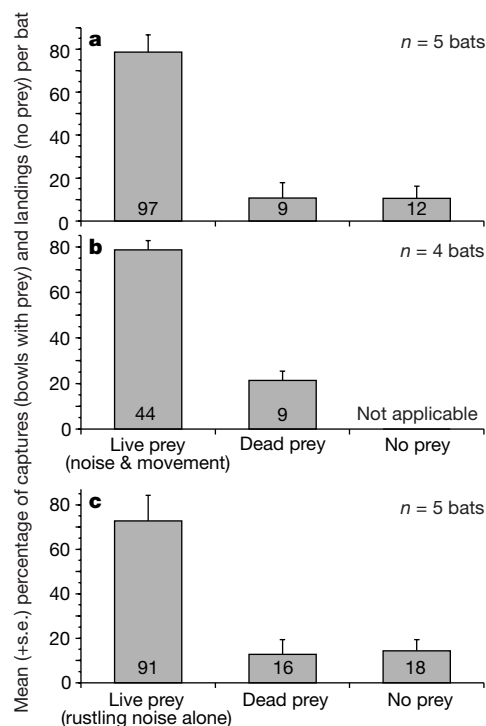


Figure 1 Experiments on foraging cues for experiments 1–3. See Methods. Data are presented as mean percentage of captures per bat (*n* = 3 *M. myotis* and 2 *M. blythii* bats; interspecific differences not significant); the standard error, s.e., indicates the inter-individual variance. The absolute number of captures (bowls with prey) or landings (no prey) is given at column foot (see Methods). **a**, Live prey generating both movement and noise, versus dead prey and no prey, under infrared illumination (two experiments per individual). **b**, Live versus dead prey in total darkness (no control in the absence of an infrared light; one experiment per individual). **c**, Live (noise-generating) prey hidden so as to render movement undetectable by echolocation, versus two controls (dead prey and no prey; each bat was submitted to this arrangement, over several trials if necessary, until 25 prey captures per individual were achieved) (infrared illumination).

frequently moving, noisy prey were captured in comparison with dead prey. Mouse-eared bats foraging in clutter were much more proficient at capturing live, that is, moving and/or noisy prey, than dead prey (Fig. 1a). In a second experiment, we tested whether infrared illumination might influence the ability to detect prey. Success in capturing live prey was unimpaired in the absence of the infrared illumination necessary for video monitoring of bats' behaviour (Fig. 1b; distributions in Fig. 1a and b show no statistical difference, $\chi^2 = 2.54$, degrees of freedom, d.f. = 2, nonsignificant), demonstrating that mouse-eared bats do not rely on vision for prey detection. In a third experiment, we tested whether bats detected the rustling noise of live locusts crawling among dead leaves, or detected prey movements through echolocation. If passive listening is the dominant detection strategy when foraging in clutter, then the patterns of selection should be similar to those found in the previous experiments. If echolocation is used for prey capture, then there should be an equivalent number of prey captures for each situation. Bats easily found hidden rustling prey that were

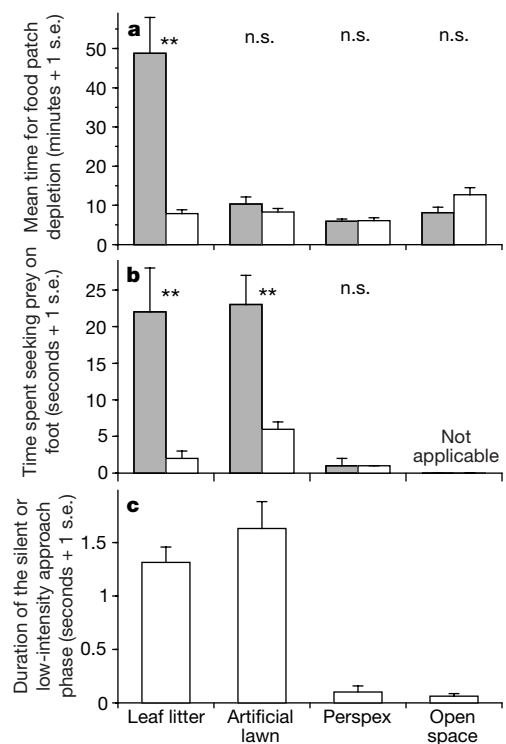


Figure 2 Experiments on foraging performance (**a**, **b**, under infrared illumination), and recordings of echolocation calls (**c**). Data were obtained from 4 *M. myotis* and 2 *M. blythii* bats. Data are presented as (**a**) average time (minutes) needed by a bat to deplete a patch containing five prey items, and (**b**) mean time (seconds) spent on foot searching for prey in the feeding arena (see Table 1 for further statistical details); s.e. represents the interindividual variance (each bat was offered three times dead and three times live locusts, and succession was randomized). Grey bars, dead prey; white bars, live prey. Probabilities are indicated by asterisks from *t*-tests performed with respect to prey status (dead versus live) for each experimental situation separately (double asterisk, $P < 0.01$; n.s., nonsignificant, **a** and **b**). In **a**, for dead prey, Tukey *post hoc* pairwise comparisons among situations yielded significantly greater values (that is, poorer performance) for leaf litter, compared with the other situations which did not differ from each other; in the case of live prey, performance was significantly poorer for open space than for leaf litter, lawn and perspex. In **b**, for dead prey, *post hoc* comparisons showed higher values (that is, poor performance) for both leaf litter and artificial lawn, compared with perspex. For live prey, artificial lawn gave a poorer performance than leaf litter and perspex. **c**, Duration (seconds) of the silent phase or low-intensity echolocation phase during approach of live prey in the same four experimental situations. *Post hoc* tests showed that the duration of the phase was significantly longer for leaf litter and artificial lawn (which did not differ from each other) compared with perspex and open space.

undetectable by echolocation (Fig. 1c): therefore, passive listening to prey-generated sounds, and not detection of prey movement by echolocation, is the basic tactic for foraging in clutter. This is because the bats' success in capturing live prey did not differ between the first and third experiment ($\chi^2 = 3.15$, d.f. = 2, non-significant).

When mouse-eared bats were required to detect silent, non-moving (dead) prey in maximum clutter (here represented by leaf litter) the average time needed to deplete the food patch was very much longer than for live prey (Fig. 2a). Moreover, when taking dead prey items in leaf litter, bats landed at random on the feeding arena and looked for prey by crawling around among dead leaves (Fig. 2b). Even a much simpler background such as an artificial lawn, with locusts protruding from the surface, yielded comparable results (Fig. 2b). Although time for patch depletion in these conditions was much faster than for leaf litter (Fig. 2a), presumably because the decision of a prospecting, flying bat to land was facilitated by the simple environmental structure, bats presented with dead locusts again landed on the feeding arena and sought prey by crawling (Fig. 2b). This suggests that the structure of the lawn was rendering the bats 'acoustically blind'. In contrast, mouse-eared bats performed much better when capturing both dead and moving locusts from the plexiglas surface, or in open space (Fig. 2a, b). Yet, the best performance overall was achieved either when bats gleaned moving, noisy prey in clutter, or when they caught prey of either status on a flat perspex surface, which mimics an acoustic mirror.

Bats produced feeding buzzes when localizing prey in air or on flat surfaces, but emitted no calls or only very faint calls at low repetition rate ('silent or low-intensity approach phase') for, on average, more than a second before detecting prey on complex surfaces (Fig. 2c). We suggest that the bats listen for prey-generated sounds during this phase. Accordingly, the low-intensity calls (optional) emitted during prey approach may detect the immediate surroundings so the bats avoid colliding with obstacles or the ground. Interestingly, a brief but intense buzz occurs immediately before landing, presumably allowing the bat to detect the ground (sound sequence on <http://www.bio.bris.ac.uk/research/bats/myotis.htm>): this provides further evidence that FM echolocation is not suited for prey detection in clutter (Table 1).

This study demonstrates first that mouse-eared bats rely on passive listening to prey-generated sounds, and not on echolocation, when foraging in echo-cluttering environments; second, that echolocation is of no use for detecting a non-moving, silent prey from cluttered backgrounds²⁰. This is especially surprising as regards the artificial lawn which represents an apparently simple substrate. These findings contradict the view that the image conveyed by the echoes of FM calls can provide detailed information about the fine texture of objects, which would allow substrate-gleaning bats to distinguish prey from cluttered surroundings^{9,24}.

Table 1 Foraging performance and echolocation

Source of variation	Sum of squares	d.f.	Variance	F ratio	P
Time for food patch depletion					
Experimental situation	3633.7	3	1211.2	17.1	<0.001
Prey status (dead versus live)	1094.6	1	1094.6	15.4	<0.001
Situation $\times$ prey status	3995.4	3	1331.8	18.8	<0.001
Error	2835.0	40	70.9		
Time spent looking for prey on foot					
Experimental situation	0.572	3	0.191	17.721	<0.001
Prey status (dead versus live)	0.288	1	0.288	26.768	<0.001
Situation $\times$ prey status	0.286	3	0.095	8.880	<0.001
Error	0.430	40	0.011		
Duration of low-intensity phase					
Experimental situation	42.63	3	14.21	32.1	<0.001
Error	31.47	71	0.443		

Analysis of variance (ANOVA) carried out on foraging performance (mean time needed for depleting a food patch, Fig. 2a; time spent looking for prey on the surface of the arena, Fig. 2b), and duration of the phase of low-intensity echolocation calls during prey approach (Fig. 2c). As species was a nonsignificant factor in a preliminary analysis, the data from the two species were grouped together in all three cases. d.f., degrees of freedom; F, F statistics; P, probability.

Assuming that evolution has prompted mouse-eared bats to adopt the most efficient acoustic compromise^{25,26}, within the framework of their sensory system, there seem to be serious limitations to FM echolocation for prey detection in echo-cluttering habitats, apparently owing to the complexity of the overlapping reflected echoes. Under such circumstances, FM echolocation rendered the bats acoustically blind¹³. This suggests that FM echolocation is mainly adapted to orientation and capture of prey either in the open space or from simple backgrounds. □

Methods

Foraging cues

In experiment 1, we placed 15 plastic bowls, 30 cm in diameter and 10 cm deep, on the floor of the lightless flight room, spreading them regularly (3×5) over an area of about 6 m². The bottom of every bowl was covered with a 2–3-cm layer of leaf litter. One-third of the bowls contained five living—that is, moving and therefore sound-generating—locusts each; another five bowls contained five dead locusts each; and the remaining five bowls received no prey (control). The arrangement of the three types of prey availability on the floor of the flight room was randomized before each experiment. Individual bats were deprived of food during the preceding 24 h and released, one at a time, for 3 h into the experimental enclosure. Foraging behaviour was recorded using an infrared video and infrared lighting, which was the sole source of light. A landing into a bowl containing prey items always resulted in the capture of one prey item.

In experiment 2, we used the same design as above, but this time the bats had to seek food in total darkness (no infrared light). We counted the prey items left in the buckets at the end of the 3 h experiments. The proportions of dead versus live prey captured in the first and second experiments were compared.

Experiment 3 was similar to the previous ones, but the possibility of detecting prey movement by echolocation was removed: leaf litter was placed in all bowls, one third of which received two live, moving locusts each, one third two dead locusts each (first control), and one third no prey (second control), but this time all bowls were entirely covered with plastic mesh, on which was placed a second, thin layer of dead leaves and a dead locust. Only the latter was available to a foraging bat. A captured locust was replaced immediately, during the course of an experiment, whilst the bat was busy eating its prey from a perch.

Foraging performance and habitat clutter

We measured the foraging performance of bats, that is, the average time—the sum of the times elapsed from the beginning of the experiment until each prey capture, divided by the number of prey captured during the trial—needed to deplete a food patch, with respect both to the status of prey (dead versus live: that is, silent versus noisy prey) and the degree of environmental clutter (four artificial micro-habitats). Feeding arenas had similar dimensions (70 $\times$ 100 cm) in the first three experimental situations. Locusts were placed at random in the arena. Each bat was offered three times five dead locusts and three times five live locusts (six experiments per individual, randomized succession).

In experiment 4, prey were suspended by a thread 40 cm below a 70 $\times$ 100 cm wire structure hanging from the flight room ceiling; in order to imitate prey moving in the air (live prey), the frame was gently swung from outside the cage using a string. Experiments lasted for up to 100 min, but were interrupted if patch depletion occurred earlier. Analyses were performed using the average individual values obtained from three trials.

Acoustics

Bats were recorded capturing live prey. We used two bat detectors (S-25, Ultra Sound Advice) placed on each side of the feeding arena along the axis of the bats' flight path and recorded echolocation calls onto two channels of a high-speed tape recorder (Racal) at 76.2 cm s⁻¹. During the approach to prey, a light signal was generated towards the infrared video camera using a photographic flash beam covered with a blacklight filter. The flash (<2 ms) enabled synchronization of sound and flight sequences. Sound digitizing (sampling rate of 44.1 kHz; input speed divided by 16) and analysis of the duration of the low-intensity echolocation phase during prey approach (defined as the part of a sequence when call intensity suddenly drops; see Fig. 2c) were obtained by using Canary 1.2.4 (Cornell Bioacoustics Workstation).

Received 13 June; accepted 2 October 2001.

- Neuweiler, G. Foraging ecology and audition in echolocating bats. *Trends Ecol. Evol.* **4**, 160–166 (1989).
- Simmons, J. A., Kick, S. A., Moffat, A. J. M., Masters, W. M. & Kon, D. Clutter interference along the target range axis in the echolocating bat, *Eptesicus fuscus*. *J. Acoust. Soc. Am.* **84**, 551–559 (1988).
- Neuweiler, G. Auditory adaptations for prey capture in echolocating bats. *Physiol. Rev.* **70**, 615–640 (1990).
- Schnitzler, H. U. & Kalko, E. K. V. Echolocation by insect-eating bats. *Bioscience* **51**, 567–569 (2001).
- Fuzessery, Z. M., Buttenhof, P., Andrews, B. & Kennedy, J. M. Passive sound localization of prey by the pallid bat (*Antrozous p. pallidus*). *J. Comp. Physiol. A* **171**, 767–777 (1993).
- Bell, G. P. Behavioral and ecological aspects of gleaning by a desert insectivorous bat, *Antrozous pallidus* (Chiroptera: Vespertilionidae). *Behav. Ecol. Sociobiol.* **10**, 217–223 (1982).
- Faure, P. A. & Barclay, R. M. R. The sensory basis of prey detection by the long-eared bat, *Myotis evotis*, and the consequences for prey selection. *Anim. Behav.* **44**, 31–39 (1992).
- Simmons, J. A. & Stein, R. A. Acoustic imaging in bat sonar: echolocation signals and the evolution of echolocation. *J. Comp. Physiol. A* **135**, 61–84 (1980).

9. Schmidt, S. Evidence for a spectral basis of texture perception in bat sonar. *Nature* **331**, 617–619 (1988).
10. Faure, P. A., Fullard, J. H. & Barclay, R. M. R. The response of tympanate moths to the echolocation calls of a substrate gleaner bat, *Myotis evotis*. *J. Comp. Physiol. A* **166**, 843–849 (1990).
11. Anderson, E. & Racey, P. A. Feeding behavior of captive brown long-eared bats *Plecotus auritus*. *Anim. Behav.* **42**, 489–493 (1991).
12. Anderson, E. & Racey, P. A. Discrimination between fluttering and non-fluttering moths by brown long-eared bats, *Plecotus auritus*. *Anim. Behav.* **46**, 1151–1155 (1993).
13. Barclay, R. M. R. & Brigham, R. M. Constraints on optimal foraging: a field test of prey discrimination by echolocating insectivorous bats. *Anim. Behav.* **48**, 1013–1021 (1994).
14. Faure, P. A. & Barclay, R. M. R. Substrate-gleaning versus aerial-hawking: plasticity in the foraging and echolocation behaviour of the long-eared bat, *Myotis evotis*. *J. Comp. Physiol. A* **174**, 651–660 (1994).
15. Arlettaz, R. Feeding behaviour and foraging strategy of free-living mouse-eared bats *Myotis myotis* and *Myotis blythii*. *Anim. Behav.* **51**, 1–11 (1996).
16. Johnston, D. S. & Fenton, M. B. Individual and population-level variability in diets of pallid bats (*Antrozous pallidus*). *J. Mammal.* **82**, 362–373 (2001).
17. Schumm, A., Krull, D. & Neuweiler, G. Echolocation in the notch-eared bat, *Myotis emarginatus*. *Behav. Ecol. Sociobiol.* **28**, 255–261 (1991).
18. Schmidt, S., Hanke, S. & Pillat, J. The role of echolocation in the hunting of terrestrial prey—new evidence for an underestimated strategy in the gleaner bat *Megaderma lyra*. *J. Comp. Physiol. A* **186**, 975–988 (2000).
19. Siemers, B. & Schnitzler, H. U. Natterer's bat (*Myotis nattereri*, Kuhl, 1818) hawks for prey close to vegetation using echolocation signals of very broad bandwidth. *Behav. Ecol. Sociobiol.* **47**, 400–412 (2000).
20. Jensen, M. E., Miller, L. A. & Rydell, J. Detection of prey in a cluttered environment by the northern bat *Eptesicus nilsonii*. *J. Exp. Biol.* **204**, 199–208 (2001).
21. Arlettaz, R., Perrin, N. & Hausser, J. Trophic resource partitioning and competition between the two sibling bat species *Myotis myotis* and *Myotis blythii*. *J. Anim. Ecol.* **66**, 897–911 (1997).
22. Arlettaz, R. Habitat selection as a major resource partitioning mechanism between the two sympatric sibling bat species *Myotis myotis* and *Myotis blythii*. *J. Anim. Ecol.* **68**, 460–471 (1999).
23. Norman, A., Jones, G. & Arlettaz, R. Noctuid moths show neural and behavioural responses to sounds made by some bat-marking rings. *Anim. Behav.* **57**, 829–835 (1999).
24. Habersetzer, J. & Vogler, B. Discrimination of surface-structured targets by the echolocating bat *Myotis myotis* during flight. *J. Comp. Physiol. A* **152**, 275–282 (1983).
25. Weibel, E. R., Taylor, C. R. & Bolis, L. *Principles of Animal Design: The Optimization and Symmorphosis Debate* (Cambridge Univ. Press, Cambridge, 1998).
26. Brigham, R. M. & Barclay, R. M. R. Prey detection by common nighthawks: does vision impose a constraint? *Ecoscience* **2**, 276–279 (1994).

Acknowledgements

We thank L. Delaloye, S. Parsons and K. Zbinden for assistance. R.A. was supported by a postdoctoral fellowship from the Swiss National Science Foundation, a research grant from the Association for the Study of Animal Behaviour, and a grant from the Vaud Academic Society. G.J. was funded by a Royal Society University Research Fellowship. We thank M. B. Fenton and J. Rydell for comments.

Correspondence and requests for materials should be addressed to R.A. (e-mail: raphael.arlettaz@nat.unibe.ch).

Evidence for the evolution of multiple genomes in arbuscular mycorrhizal fungi

Gerrit Kuhn^{*†}, Mohamed Hijri^{*†} & Ian R. Sanders^{*}

^{*} Institute of Ecology, University of Lausanne, Biology Building, 1015 Lausanne, Switzerland

[†] These authors contributed equally to this work

Ancient asexuals directly contradict the evolutionary theories that explain why organisms should evolve a sexual life history^{1,2}. The mutualistic, arbuscular mycorrhizal fungi are thought to have been asexual for approximately 400 million years^{3,4}. In the absence of sex, highly divergent descendants of formerly allelic nucleotide sequences are thought to evolve in a genome². In mycorrhizal fungi, where individual offspring receive hundreds of nuclei from the parent, it has been hypothesized that a population of genetically different nuclei should evolve within one individual^{5,6}. Here we use DNA–DNA fluorescent *in situ* hybridization to show that genetically different nuclei co-exist in individual arbuscular

mycorrhizal fungi. We also show that the population genetics techniques⁴ used in other organisms are unsuitable for detecting recombination because the assumptions and underlying processes do not fit the fungal genomic structure shown here. Instead we used a phylogenetic approach to show that the within-individual genetic variation that occurs in arbuscular mycorrhizal fungi probably evolved through accumulation of mutations in an essentially clonal genome, with some infrequent recombination events. We conclude that mycorrhizal fungi have evolved to be multi-genomic.

Arbuscular mycorrhizal fungi (Class Zygomycetes; Order Glomales) are extremely successful fungi that form mutualistic symbioses with the roots of approximately 60% of all plant species⁷. They improve plant nutrition and promote plant diversity⁸. These fungi have been assumed to be asexual⁷. This is supported by measurements of the degree of linkage disequilibrium, which indicated that genetic variation among the spores of arbuscular mycorrhizal fungi deviates significantly from that expected from a recombinant population⁴. Genetic diversity in the ribosomal DNA occurs inside individual spores^{9–12}, even though it is thought that several copies of rDNA are kept the same by concerted evolution¹³. It has been hypothesized that by accumulation of mutations, in the absence of recombination, individual arbuscular mycorrhizal fungi have evolved to comprise genetically divergent nuclei, or that one individual contains several genomes³. Here we refer to an arbuscular mycorrhizal fungal spore as an individual.

We tested the hypothesis that individuals contain genetically different nuclei by performing specific fluorescent DNA–DNA *in situ* hybridization (FISH) on nuclei from spores of the arbuscular mycorrhizal fungus *Scutellospora castanea* (BEG 1). We used hybridization probes that specifically recognize two divergent sequences of the ITS2 region, known as T2 and T4, that were previously shown to co-occur within individual spores of this fungus¹⁴. Probes were only used for variant ITS2 sequences that had previously been shown to

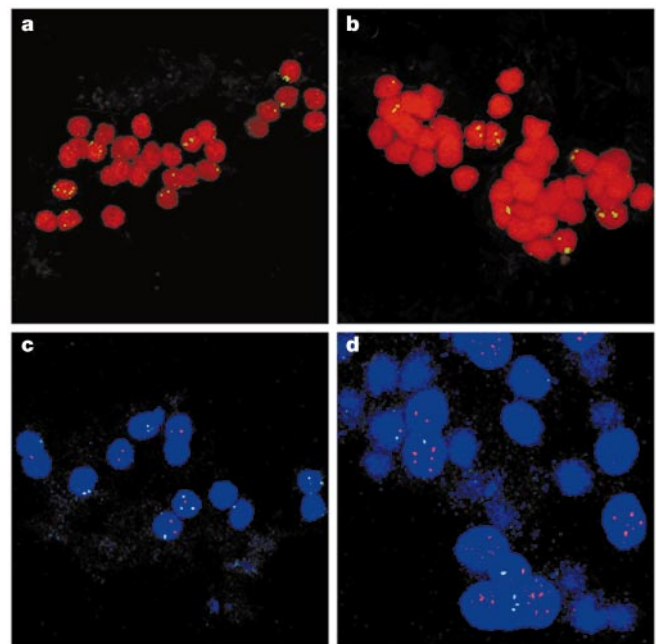


Figure 1 Nuclei of *Scutellospora castanea* taken with scanning laser confocal microscopy after single-target and double-target DNA–DNA FISH. **a**, Hybridization signals (green) of the probe T2-DIG to nuclei (red). **b**, Hybridization signals (green) of the probe T4-DIG to nuclei (red). **c**, Hybridization signals of the probes T2-DIG (light blue) and T4-biotin (red) to nuclei (purple). **d**, Hybridization signals of the probes T2-biotin (red) and T4-DIG (light blue) to nuclei (purple). The colours of images **c** and **d** have been adjusted to give better contrast between the colour of the two probes and the nuclei.

Table 1 DNA–DNA FISH on *S. castanea* nuclei

Treatments	Number of nuclei observed	% nuclei hybridizing with T2	% nuclei hybridizing with T4	% nuclei hybridizing with T2 and T4	% nuclei unlabelled
Single FISH with T2 (digoxigenin)	1,662	40.26 (3.04)	—	—	59.74 (3.04)
Single FISH with T4 (digoxigenin)	3,405	—	17.03 (2.77)	—	82.97 (2.77)
Double FISH with T2 (biotin) and T4 (digoxigenin)	1,120	40.16 (2.87)	6.52 (1.04)	8.20 (2.30)	45.12 (3.46)
Double FISH with T2 (digoxigenin) and T4 (biotin)	1,364	41.02 (3.38)	9.64 (2.48)	9.63 (2.39)	39.71 (4.16)

Numbers in parentheses represent $\pm$ 1 s.e.m. Five spores were crushed on each slide and the number of sides observed were 34, 43, 28 and 23 for the treatments single-target FISH (T2-DIG), single-target FISH (T4-DIG), double-target FISH (T2-biotin and T4-DIG) and double-target FISH (T2-DIG and T4-biotin), respectively. Proportions of nuclei that were labelled with T2 or T4 did not differ from the presented results in experiments with a smaller sample size, where one spore was placed on each slide. Student's *t*-tests showed significant differences between the percentage of nuclei labelled with T2 and T4 in single-target FISH ($t_{1,75} = 30.31, P \leq 0.0001$) and between T2 and T4 in both double-target FISH experiments ($t_{1,44} = 121.76, P \leq 0.001$ and $t_{1,54} = 56.05, P \leq 0.0001$). The *t*-tests revealed no significant differences in the percentage of nuclei hybridizing to either T2 or to T4 as a result of different labelling and also showed no significant differences in hybridization percentage with a given probe in single-target FISH experiments compared to double-target FISH experiments.

be of glomalean origin¹⁵. Single-target FISH showed that significantly more nuclei in *S. castanea* spores contained sequence T2 (40%) than T4 (17%) (Fig. 1a, b; Table 1). Double-target FISH using T2 and T4 probes showed that the divergent sequences T2 and T4 were indeed segregated in different frequencies among the nuclei (Fig. 1c, d; Table 1). Approximately 40% of nuclei contained only the T2 sequence and between 6 and 9% of nuclei contained only the T4 sequence. T2 and T4 co-occurred in between 8 and 9% of nuclei (Fig. 1c, d; Table 1). These results support the hypothesis that arbuscular mycorrhizal fungal spores contain a population of genetically different nuclei.

We need to know whether this genetic variation has been brought about by lack of recombination. Because several genomes exist within individuals, recombination could potentially occur among nuclei within individuals and this has not previously been considered. Using a theoretical approach we assessed whether the prediction from previous population genetic studies that mycorrhizal fungi are clonal (based on fingerprinting and calculations of index of association⁴) is valid. We constructed artificial data sets representing presence or absence of alleles at 15 loci for a population of recombining nuclei within 30 spores. We then calculated the total

number of nuclei containing an allele at each of the loci, for each spore. However, a fingerprinting technique performed on whole spores cannot take into account within-individual polymorphism (Fig. 2a). Therefore, whether the presence or absence of an allele at a given locus would be observed for each spore would depend on the number of nuclei in which that allele is present and the sensitivity of the fingerprinting technique (Fig. 2b). We therefore calculated the presence of alleles at each locus for each spore, simulating different levels of sensitivity of the fingerprinting technique. We then calculated the index of association¹⁶ for the population spores. All observed values of the index of association deviated significantly from zero (Fig. 3), allowing us to reject the null hypothesis that the population of hypothetical arbuscular mycorrhizal fungal spores were a recombining population. We concluded that any potential recombination was hidden within individuals and was not detectable using this technique.

Thus, the genomic structure of arbuscular mycorrhizal fungi gives rise to difficulties in detecting recombination with techniques that are based on linkage disequilibrium. To solve this problem, we used a phylogenetic technique known as character incompatibility analysis¹⁷ to detect whether genetically divergent nuclei in these fungi are likely to have arisen by the accumulation of mutations in clonal nuclear lineages or by recombination events. We looked at sequence variation in ITS regions (including the 5.8S gene) within isolates of the arbuscular mycorrhizal fungi *Glomus geosporum*, *Glomus mosseae* and *Gigaspora margarita* and in the 28S gene in *G. geosporum*, *Glomus coronatum*, *Glomus constrictum* and *G. mosseae*¹⁸ (Table 2). Calculation of the Le Quesne probability¹⁷ showed that for a large proportion of the variable characters the incompatibility count differed significantly from that which would be expected if it had arisen from recombination. A small proportion of the variation in rDNA sequences could be explained by recombination events, although this was higher in *Gi. margarita* than for the other fungi. Using a jackknife procedure we showed that only a small proportion of the sequences contributed nearly all the variation that is explained as being the result of recombination events (Table 2). This was also true for the variation seen in *Gi. margarita*. The results show, therefore, that most of the variant sequences are the result of accumulation of mutations in a clonal genome.

Until now, reports of within-individual sequence variation in arbuscular mycorrhizal fungi have been restricted to rDNA sequences. Our results demonstrate genetic differences only among nuclei for one region of rDNA. Therefore, we also provide evidence to support genetic differences among fungal nuclei, clonality and Muller's ratchet. We analysed the sequence of part of a gene encoding for a binding protein (*BiP* gene) that has high amino-acid sequence similarity (86% and 78%) to that of *Aspergillus niger* and *Saccharomyces cerevisiae*, respectively¹⁹. This gene is highly conserved in eukaryotes and is a single-copy gene in other fungi. We sequenced 15 variant sequences of this gene from genomic DNA of one isolate of *Glomus intraradices*. Along a length

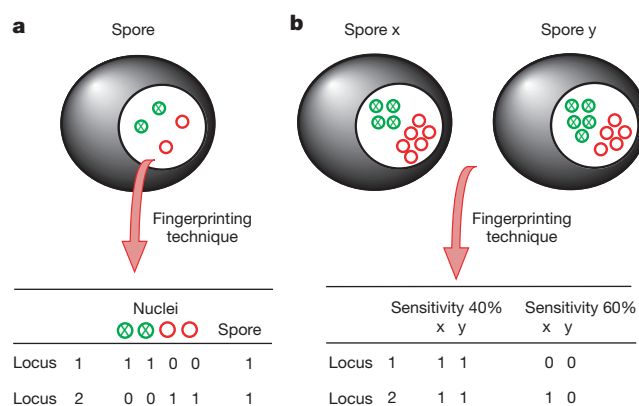


Figure 2 Within- and among-spore polymorphism. **a**, A spore of a hypothetical arbuscular mycorrhizal fungus contains a population of two genetically different nuclei (crossed and open circles). Within-individual polymorphism caused by differences among nuclei is masked in a fingerprinting method on the whole spore because the presence of both alleles is detected. **b**, Two hypothetical spores of an arbuscular mycorrhizal fungus, x and y, contain a population of two genetically different nuclei that are present in different frequencies. At a sensitivity of fingerprinting of 40%, where at least four nuclei carrying a given allele are required to show its presence in the spore, spores x and y appear to be genetically identical. At sensitivity 60%, where six nuclei are required to show the presence of an allele, spores x and y appear to be genetically different. Thus, spores can appear to be polymorphic or identical, depending on the fingerprinting sensitivity. Binary matrices show presence of alleles (1) or absence (0) during detection using a fingerprinting method at two different sensitivities.

Table 2 Incompatibility analysis on the variation in rDNA sequences

Species	Number of variable sequences	Number of clonal characters (% of variant characters)	Number of recombinant characters (% of variant characters)	MIC 50†	MIC 10‡
Analysis performed on variable ITS sequences (comprising ITS1, 5.8S gene and ITS2)					
<i>G. geosporum</i> *	26	50 (83%)	10 (17%)	3%	7%
<i>G. geosporum</i> †	26	17 (85%)	3 (15%)	—	—
<i>G. mosseae</i> *	22	102 (68%)	47 (32%)	13%	54%
<i>G. mosseae</i> †	22	47 (70%)	20 (30%)	—	—
<i>Gi. margarita</i> *	18	53 (52%)	48 (48%)	11%	33%
<i>Gi. margarita</i> †	18	10 (31%)	22 (69%)	—	—
Analysis performed on variable sequence of 28S gene					
<i>G. geosporum</i> *	15	122 (92%)	10 (8%)	13%	26%
<i>G. coronatum</i> *	19	80 (86%)	13 (14%)	15%	21%
<i>G. constrictum</i> *	12	21 (51%)	20 (49%)	8%	8%
<i>G. mosseae</i> *	17	57 (90%)	5 (8%)	5%	10%

* The analysis was performed with insertions or deletions in the data set representing a fifth character.

† The analysis was performed on sequences that varied only due to substitution.

‡ MIC 50 and MIC 10 are defined as the percentage of sequence variants that need to be removed to reduce the matrix incompatibility to less than 50% and less than 10% of the total matrix incompatibility in the data set. This is performed by sequentially eliminating the sequences that contribute the most to the total matrix incompatibility and then recalculating the total incompatibility contained in the remaining data set.

of 680 base pairs (bp), these variants ranged in similarity at the nucleotide level from 92% to 99% from one of the sequences that was randomly chosen as the comparison sequence (see the Supplementary Information). If this gene is also single-copy in arbuscular mycorrhizal fungi then the sequence variants that we observed must be segregated among nuclei. Incompatibility analysis indicated that

no variation in these sequences was likely to be due to recombination. Furthermore, both synonymous and non-synonymous substitutions occurred in the sequences; that is, not all of the substitutions are selectively neutral. The mean number of substitutions compared to the comparison sequence was 26.6 (range 3 to 48) with a mean ratio of 1.02 (s.e.m. $\pm$ 0.10) synonymous to one non-synonymous substitution. We would expect a higher rate of synonymous substitutions to non-synonymous substitutions unless Muller's ratchet is in operation, because selection should act to conserve the sequence at non-synonymous sites. These data are consistent with evidence for Muller's ratchet in genomes of endosymbiotic prokaryotes²⁰.

From our results, we predict that genetic variation is generated by accumulation of mutations in a predominantly clonal genome, leading to the creation of a population of genetically different nuclei. The evidence for (infrequent) recombination events is unsurprising in an organism that contains genetically different nuclei that co-exist coenocytically. However, our analyses indicate that these recombination events are rare and do not purge the majority of mutations in the genomes. Many other fungi have stages of heterokaryosis, where more than one nuclear genotype coexists, although genetic bottlenecks occur in the life cycle of each individual at each generation, limiting the number of nuclei that are transferred to the next generation to one per individual. No such stage in the life history of arbuscular mycorrhizal fungi is known and new spores receive many nuclei from the mother hyphae. Spores may also receive genetic material by fusion of hyphae^{21,22}, although the studies suggest that cross-incompatibility is most frequent. We suggest, therefore, that such a mechanism would not reduce the potential problems associated with Muller's ratchet as the nuclei received would also be accumulating mutations. Most theories of evolutionary and population genetics assume that one individual contains a single genome and so new models for evolution of multi-genomic organisms need to be developed. Our results may stimulate research on mycorrhizal symbiosis at the molecular and physiological levels, because several variants of the same gene exist in an individual and could lend to variation in expression. Our results provide essential information that is needed before any studies on genomics or proteomics of arbuscular mycorrhizal fungi are undertaken. □

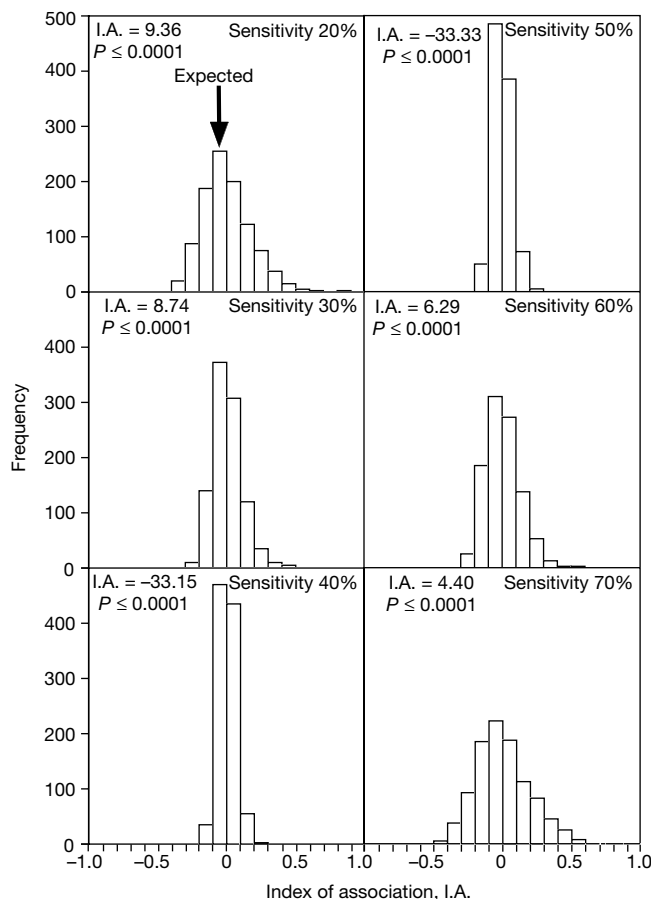


Figure 3 Frequency distributions for the index of association for a population of 30 hypothetical spores of an arbuscular mycorrhizal fungus. Each spore contains a population of recombining nuclei, calculated at 20% to 70% sensitivities of the fingerprinting. Data calculated with 80% sensitivity did not give a normal frequency distribution and is not shown. Each graph also shows the observed I.A. for the data set: in each case this differed significantly from the expected I.A.

Methods

Slot-blot and DNA–DNA FISH

Hybridization probes were constructed for two different sequences of the ITS2 region, subsequently referred to as T2 and T4, that were known to be variable in *S. castanea* (GenBank accession numbers SCAJ2872 and SCAJ2874)^{14,15}. Probes were produced from cloned DNA that was amplified with the forward primer (5'-CACCTGCTTGAGGGT-

CAGT-3') and the reverse primer ITS4 (ref. 23). The size of the probes were 274 bp and 258 bp for T2 and T4, respectively. Probes were labelled by polymerase chain reaction (PCR) with DIG-11-dUTP and Biotin-16-dUTP (Roche) and purified with QiaQuick PCR purification kit (Qiagen).

Spores of *S. castanea*, *G. geosporum* (BEG 18) and *Glomus* sp. (BEG 19) were collected from pot cultures and immediately fixed (in 4% formaldehyde, 100 mM Tris/HCl, pH 8, 100 mM NaCl, 2 mM MgCl₂ and 0.05% Triton X100, for 2 h at room temperature). Five spores or single spores were crushed on Frost Plus slides (Polylabo) and then dried overnight. For single-target FISH, 25 µl of the hybridization solution was loaded per slide. For double-target FISH, a mixture of digoxigenin and biotin probes in a 1:1 ratio was added to the hybridization solution. Hybridization was carried out at 37 °C overnight. The post-hybridization washes were made twice with 50% formamide in double-strength SSC (42 °C, 10 min), then twice with double-strength SSC, (37 °C, 5 min) and then rinsed again twice with half-strength SSC with 0.1% SDS (60 °C, 15 min). The post-hybridization washes were identical to those used for the slot-blot procedure. For full details regarding slot-blot conditions and slide preparation for FISH and all controls see the Supplementary Information.

For signal detection in single-target FISH, slides were incubated with anti-DIG-fluorescein conjugate antibody (Roche) and counterstained with propidium iodide. In double-target FISH experiments, signals were detected with a mixture of anti-DIG-fluorescein conjugate antibody (Roche) and Streptavidin Texas-Red conjugate. The slides were counterstained with TOTO-3 (molecular Probes). Slides were examined using a scanning confocal microscope (further details are given on signal detection and microscopy in the Supplementary Information).

Testing for detection of recombination

Thirty identical data sets were constructed, representing 30 spores of arbuscular mycorrhizal fungi, each containing a population of 50 genetically different nuclei that were variable at 15 loci. The nuclei within each spore were then recombined by the random rearrangement of alleles among nuclei, assuming that loci were independent of each other. The frequency of genetically different nuclei occurring in each spore was then altered by randomly selecting 10 nuclear genotypes that were then replicated a random number of times in the spore, with a maximum limit of 10 replicates of a given nuclear genotype per spore. The number of nuclei showing the presence of an allele at each locus was summed for each spore. The presence or absence of an allele at each locus was then calculated for each whole spore using different sensitivities of the fingerprinting method to give binary data sets. Sensitivity was defined by the percentage of nuclei containing a given allele that was required to give a positive signal with a fingerprinting technique. The sensitivities ranged from 100% (where the presence of an allele in 1 nucleus in a spore was sufficient to give a positive signal with a fingerprinting technique) through to 0% (where even if every nucleus contained the allele the technique would not be sensitive enough to detect presence of the allele). Polymorphism among the 30 spores at the 15 loci was calculated for data sets at 20%, 30%, 40%, 50%, 60%, 70% and 80% sensitivities. Sensitivity was calculated on a percentage basis rather than actual numbers of nuclei to account for differences in the numbers of nuclei among spores. Each of these seven data sets were then separately used to calculate the index of association¹⁶.

Incompatibility analysis on rDNA and *BIP* gene sequences

Incompatibility analysis was performed on sequences of ITS and 28S rDNA from different arbuscular mycorrhizal fungi species. ITS sequences of *G. geosporum* (BEG 18) were obtained by extracting genomic DNA from spores (Qiagen DNeasy Plant Mini Kit) from a culture that originated from a single spore and amplified by PCR². The 640-bp product containing ITS1, the 5.8S gene and ITS2 was purified (Qiagen QiaQuick purification kit), ligated into a pGEM-T vector and transformed into *Escherichia coli* JM109 (Promega). Clones were selected at random and both strands were sequenced using a dye-terminator cycle sequencing kit. Sequences were carefully edited by hand. ITS sequences from *G. geosporum* (BEG 18) and from *G. mosseae*²⁵ and *G. margarita*²⁴ were aligned. Sequences of the 28S gene for the species *G. geosporum* (BEG 11)¹⁸, *G. coronatum* (BEG 49)¹⁸, *G. constrictum* (BEG 130)¹⁸ and *G. mosseae* (BEG 25)¹⁸ were also aligned and used for the analysis. All isolates originated from single spores except *G. coronatum* and *G. constrictum*. Further details on the compatibility analysis on rDNA and the gene encoding a binding protein are given in the Supplementary Information.

Received 20 August; accepted 4 October 2001.

1. Maynard Smith, J. Contemplating life without sex. *Nature* **324**, 300–301 (1986).
2. Welch, D. M. & Meselson, M. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288**, 1211–1215 (2000).
3. Judson, O. P. & Normark, B. B. Ancient asexual scandals. *Trends Ecol. Evol.* **11**, 41–46 (1996).
4. Rosendahl, S. & Taylor, J. W. Development of multiple genetic markers for studies of genetic variation in arbuscular mycorrhizal fungi using AFLP. *Mol. Ecol.* **6**, 821–829 (1997).
5. Sanders, I. R., Clapp, J. P. & Wiemken, A. The genetic diversity of arbuscular mycorrhizal fungi in natural ecosystems: A key to understanding the ecology and functioning of the mycorrhizal symbiosis. *New Phytol.* **133**, 123–134 (1996).
6. Sanders, I. R. No sex, we're fungi. *Nature* **399**, 737–739 (1999).
7. Smith, S. E. & Read, D. J. *The Mycorrhizal Symbiosis* (Academic, Oxford, 1997).
8. Van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
9. Sanders, I. R. *et al.* Identification of ribosomal DNA polymorphisms among and within spores of the Glomales: application to studies of genetic diversity of arbuscular mycorrhizal fungal communities. *New Phytol.* **130**, 419–427 (1995).
10. Lloyd-MacGill, S. A. *et al.* Diversity of the ribosomal internal transcribed spacers within and among isolates of *Glomus mosseae* and related mycorrhizal fungi. *New Phytol.* **133**, 103–111 (1996).

11. Redecker, D., Thierfelder, H., Walker, C. & Werner, D. Restriction analysis of PCR-amplified internal transcribed spacers of ribosomal DNA as a tool for species identification in different genera in the order glomales. *Appl. Environ. Microbiol.* **63**, 1756–1761 (1997).
12. Lanfranco, L., Delpero, M. & Bonfante, P. Intraspecific variability of ribosomal sequences in the endomycorrhizal fungus *Gigaspora margarita*. *Mol. Ecol.* **8**, 37–45 (1999).
13. Hoelzel, A. R. & Dover, G. A. *Molecular Genetic Ecology* 1–20 (IRL, Oxford, 1991).
14. Hijiri, M., Hosny, M., van Tuinen, D. & Dulieu, H. Intraspecific ITS polymorphism in *Scutellospora castanea* (Glomales, Zygomycota) is structured within multinucleate spores. *Fungal Genet. Biol.* **26**, 141–151 (1999).
15. Redecker, D., Hijiri, M., Dulieu, H. & Sanders, I. R. Phylogenetic analysis of a dataset of fungal 5.8S rDNA sequences shows that highly divergent copies of Internal Transcribed Spacers reported from *Scutellospora castanea* are of Ascomycete origin. *Fungal Genet. Biol.* **28**, 238–244 (1999).
16. Maynard Smith, J. *et al.* How clonal are bacteria? *Proc. Natl Acad. Sci. USA* **90**, 4384–4388 (1993).
17. Mes, T. H. M. Character compatibility of molecular markers to distinguish asexual and sexual reproduction. *Mol. Ecol.* **7**, 1719–1727 (1998).
18. Clapp, J. P., Rodriguez, A. & Dodd, J. C. Inter- and intra-isolate rRNA large subunit variation in *Glomus coronatum* spores. *New Phytol.* **149**, 539–554 (2001).
19. van Gemen, I. A. *et al.* The ER chaperone encoding *bipA* gene of black Aspergilli is induced by heat shock and unfolded proteins. *Gene* **198**, 43–52 (1997).
20. Moran, N. A. & Wenegreen, J. J. Lifestyle evolution in symbiotic bacteria: Insights from genomics. *Trends Ecol. Evol.* **15**, 321–326 (2000).
21. Giovannetti, M., Azzolini, D. & Citeresi, A. S. Anastomosis formation and nuclear and protoplasmic exchange in arbuscular mycorrhizal fungi. *Appl. Environ. Microbiol.* **65**, 5571–5575 (1999).
22. Giovannetti, M. *et al.* The occurrence of anastomosis formation and nuclear exchange in intact arbuscular mycorrhizal networks. *New Phytol.* **151**, 717–724 (2001).
23. White, T. J., Bruns, T., Lee, S. & Taylor, J. in *PCR Protocols: A Guide to Methods and Applications* (eds Innes, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J.) 315–322 (Academic, New York, 1990).
24. Antoniolli, Z. I., Schachtman, D. P., Ophel-Keller, K. & Smith, S. E. Variation in rDNA ITS sequences in *Glomus mosseae* and *Gigaspora margarita* spores from permanent pasture. *Myc. Res.* **104**, 708–715 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank T. Boller and A. Wiemken for allowing part of this work to be conducted in the Botanical Institute Basle, T. Mes and M. Wilkinson for advice on the analysis, A. Rodriguez and J. P. Clapp for providing 28S sequences, E. Stöckli and H. Reichert for use of the confocal microscope, P. Heslop-Harrison for advice on *in situ* controls and J. P. Clapp, M. G. A. van der Heijden, L. Keller, A. M. Koch and F. Mery for critically reading the manuscript. We thank the Swiss National Science Foundation for supporting this work with a standard research grant and a fellowship awarded to I.R.S. under the professorial fellowship programme.

Correspondence and requests for materials should be addressed to I.R.S. (e-mail: Ian.Sanders@ie-bsg.unil.ch).

Transmission potential of smallpox in contemporary populations

Raymond Gani & Steve Leach

Centre for Applied Microbiology and Research, Porton Down, Salisbury, Wiltshire SP4 0JG, UK

Despite eradication¹, smallpox still presents a risk to public health whilst laboratory stocks of virus remain^{2,3}. One factor crucial to any assessment of this risk is R_0 , the average number of secondary cases infected by each primary case. However, recently applied estimates have varied too widely (R_0 from 1.5 to >20) to be of practical use, and often appear to disregard contingent factors such as socio-economic conditions and herd immunity^{4–8}. Here we use epidemic modelling⁹ to show a more consistent derivation of R_0 . In isolated pre-twentieth century populations^{10–12} with negligible herd immunity, the numbers of cases initially rose exponentially, with an R_0 between 3.5 and 6. Before outbreak controls were applied, smallpox also demonstrated similar levels of transmission in 30 sporadic outbreaks in twentieth century Europe¹, taking into account pre-existing vaccination levels^{13,14} (about 50%) and the role of hospitals in doubling early transmission. Should smallpox recur, such estimates of transmission potential (R_0

from 3.5 to 6) predict a reasonably rapid epidemic rise before the implementation of public health interventions, because little residual herd immunity exists now that vaccination has ceased.

Before eradication, smallpox caused significant epidemics and mortality¹. Following the WHO (World Health Organization) campaign, smallpox was eradicated by 1979 and routine vaccination gradually ceased worldwide¹. Official stocks of virus now remain in only two laboratories, one at CDC in Atlanta, Georgia, and the other at VECTOR, Novosibirsk, Russia¹. There is concern, however, that the virus might be held clandestinely and less securely elsewhere³. Release of smallpox back into the population could constitute a significant public health problem on account of the transmissibility of the virus and waning immunity. The greater the transmissibility of the virus the greater its potential to spread, and the more intense are the intervention strategies needed to bring it under control. Significant discrepancies exist in the recent literature^{4–8}, so it is essential to provide a rational assessment of the transmissibility of smallpox by reference to adequate and relevant data.

Reports of outbreaks of smallpox caused by variola major were sought that contained sufficient numerical detail to derive estimates of transmissibility, R_0 (see Methods). Such reports are scarce, and those in Table 1 represent all that were found from the extensive published data. Data for four of the outbreaks were from the pre-vaccination era. Analysis of the cumulative smallpox deaths for Boston, USA¹⁰ in 1721 and Burford, England¹¹ in 1758 (Fig. 1a, b and Table 1) gave R_0 values of 4.3 and 3.4, respectively, which produced extremely good fits to the data. As was often the case for small isolated communities, nearly the whole population of Burford was susceptible. Similarly, outbreaks in Chester¹⁰ (1774) and Warrington¹⁰ (1773) (Fig. 1c, d, and Table 1) gave estimates for R_0 of 5.8 and 4.7 (range for Warrington 4–5.3), respectively. In both of these outbreaks, only children less than 10 years old were afflicted, so the numbers of susceptible individuals were estimated from the relevant birth and age-dependent mortality rates (Table 1). The predicted epidemic curve fitted the data for Chester very well. For Warrington, however, a range of R_0 values was derived that depended on the estimated population size, which was calculated to be between 6,000 and 8,000, respectively. The fitted epidemic curve that is shown for Warrington, which is based on an R_0 of 4.7 and a population estimate of 7,000, followed the initial course of the observed epidemic well but over-predicted the cumulative number of deaths by the end of the epidemic. It is possible that the case fatality rate of 25% given by Creighton¹⁰ was an overestimate as the other contemporaneous epidemics modelled here had case fatality rates of about 15%. Adjusting the case fatality rate to 15% greatly improved the fit to the data, closely fitting all 12 data points and

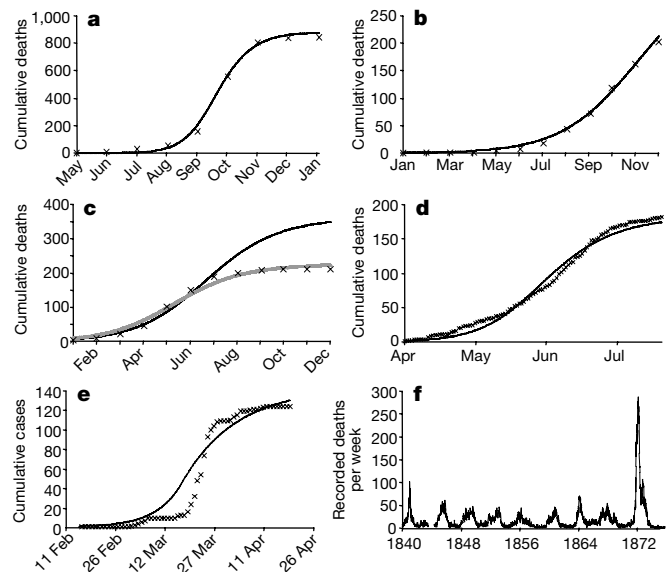


Figure 1 Analysis of smallpox epidemics. **a–e**, Epidemic curves (solid lines) fitted to data (points) from selected smallpox outbreaks. **a**, Boston monthly deaths; **b**, Chester monthly deaths; **c**, Warrington monthly deaths, predicted using published mortality rate of 25% (black line), or 15% (grey line; see text); **d**, Burford daily deaths; and **e**, Kosovo daily onset of cases. Disease and population parameters given in the text and Tables 1 and 2, respectively. **f**, Endemic smallpox in London with epidemics occurring approximately every 4 years¹², giving a lower bound for the estimate of R_0 of about 5; see text and Methods.

producing an estimated R_0 value of 4.8 (see Fig. 1d and Table 1).

Smallpox was endemic in London from the seventeenth to the twentieth century. Weekly mortality statistics were recorded, first in the London Bills of Mortality and later in the Registrar General's reports¹². Superimposed on an endemic level of disease, epidemics generated a cyclical pattern of smallpox deaths¹⁵ (Fig. 1f). Such patterns are characteristic of the dynamics of endemic diseases and the models that are often used to describe them⁹. Analysis of these individual epidemics using the methods described above was not possible because independent estimates of the numbers susceptible in the population could not be made. The poorly documented levels of variolation¹ and vaccination that were being practised compounded the problem. If the problems due to vaccination are neglected, the inter-epidemic interval¹⁵ between 1836 and 1890

Table 1 Population parameters and analysis of smallpox epidemics for estimation of R_0

Outbreak	Population size	Number susceptible	Case fatality rate (%)	Estimated R_0	Reference
Boston, USA (1721)	10,565*	6,739*	14*	4.3	10
Burford (1758)	1,520*	1,519*	13*	3.4	11
Chester (1774)	14,713*	3,167†	15*	5.8	10
Warrington (1773)	7,000 (6,000–8,000)§	2,250‡	25*	4.7 (4.0–5.3)	10
Paris, France (1766)	NA	NA	NA	>4–5¶	16
London (1836–1870)	NA	NA	NA	>5¶	12, 10, 15
Kosovo (1972)	2,200,000#	1,100,000*	10**	10.8*	1
Europe (1958–1973)	NA	NA	10**	5.4 (discounting hospital-associated cases)††	1
(Summary of 30 importations)‡‡				10 to 12	
				5 to 6 (discounting hospital-associated cases)††	

* Given in source.

† Only individuals ≤ 10 years were susceptible, number estimated from a birth rate of 0.03 (England) and age-dependent mortality rates (Chester).

‡ Only individuals ≤ 10 years were susceptible, number estimated from birth rate 0.025 and age-dependent mortality rates (Warrington).

§ Range estimated from population size of 9,501 given for 1781 adjusted for births after 1773 (Warrington's population was expanding due to industrial success).

|| A revised estimate for case fatality rate (see text).

¶ London estimated from inter-epidemic interval (R not R_0 because of vaccination) and Paris estimated from force of infection and life expectancy (see text).

1998 estimate (<http://archive.tol.cz/countries/yug98.html>), value not critical as epidemic was quickly stopped with only 125 cases.

** Estimate based on 50% vaccination coverage (see text).

†† Given in source but not critical as model was fitted to cases not deaths.

‡‡ Hospital-associated cases accounted for 50% of transmission.

‡‡ Estimate of R_0 based on the increasing number of cases per importation caused by increasing delays in outbreak notification (see text).

NA, not applicable.

Table 2 Interventions imposed for smallpox outbreak in Kosovo, 1972

Parameter	Value	Meaning	Reference
χ_1	0.06 days ⁻¹	(Contact quarantine period) ⁻¹	24
χ_2	0.04 days ⁻¹	(Infectious quarantine period) ⁻¹	1
ϵ_1	0.975	Vaccine efficacy when uninfected	13, 23
ϵ_2	0.30	Vaccine efficacy when infected and latent	1, 23
ρ	0.975	Proportion of contacts found	1
θ	0.95	Daily removal rate of infectious individuals into quarantine	1

suggests that the transmission rate, R (see Methods) was about 5. However, vaccination was being practised, so this provides a lower bound for the estimate of R_0 . Interepidemic intervals as low as 2–3 years¹⁵ and presumably higher values of R_0 had also applied previously in London, and these have been variously correlated with greater susceptibility to disease, greater density of population or lack of vaccination. Dietz and Heesterbeek¹⁶ also report on an analysis of smallpox in Paris by Bernoulli in 1766. His estimated force of infection, λ , of 0.125, and life expectancy, L , of 32 years, give an estimated R_0 of between 4 and 5 for smallpox in this other major European city (given that $R_0 = \lambda L$ or $R_0 = 1 + \lambda L$).

Although no longer endemic in Europe by the early 1950s, smallpox still caused isolated epidemics after the importation of the disease from endemic areas despite the considerable level of herd immunity in European populations^{1,13,17}. One of the best documented epidemics was in Yugoslavia in 1972 (ref. 1). Starting in Kosovo, the outbreak went unrecognized until the second generation of infection, when vaccination and quarantine of contacts was initiated, followed by vaccination of about 95% of the population. We fitted equation (1) to the data, parameterized to take into account these interventions (Table 2), and hence estimated R (see Methods) to be about 5.4; this value fell rapidly to below 1 with the introduction of quarantine and vaccination (Fig. 1e, Table 1). However, about 50% of the population were probably protected by vaccinations that pre-dated the outbreak^{13,14}, such that the initial R_0 would have been nearer 10.8. The epidemic curve from equation (1) provided a reasonable overall fit to the data from Kosovo but underestimated the number of cases in the earlier part of the outbreak, probably on account of the small size of the outbreak and the fact that there were only two generations of infection before interventions were applied.

Thirty European smallpox importations between 1958 and 1973 were also briefly summarized in ref. 1. Stratified by the time delay between importation and recognition of the outbreak, and its notification to the WHO (0–7, 8–14, 15–21 and 22–42 days), the outbreaks that took longer to be notified resulted in a greater average number of cases before the epidemics were halted (5.4, 8.8, 15.8, 67.8 cases, respectively). With the median delay adjusted by the addition of a latent period for the index case and expressed in units of intergeneration time (19 days), a clear relationship was derived between the delay and the total number of cases (Fig. 2a). Each generation's delay increased the average number of cases by $e^{1.71}$, giving a mean estimate for R of 5.5 ($\mu \pm \text{s.e.m.} = 4.9$ to 6.3). As before, these estimates are affected by previous vaccination levels; 50% vaccination, as has been suggested^{13,14}, would give an estimate for R_0 of about 11. Interestingly, about 50% of cases in most of these European outbreaks, including in Kosovo, were acquired in hospital by staff and patients¹ before the outbreak was recognized and hospital infection-control measures had been put in place (Fig. 2b). Without hospital-associated cases, the R_0 for smallpox would have been about 5.5 for community-acquired disease.

As expected from data on other infectious agents⁹, the estimates of R_0 derived for smallpox varied with time and place. Over the 1700s to 1900s, they tended to be in the range 3.5–6, but under particular circumstances rose to nearer 10–12 (that is, with crowding and poorer socio-economic conditions in eighteenth century

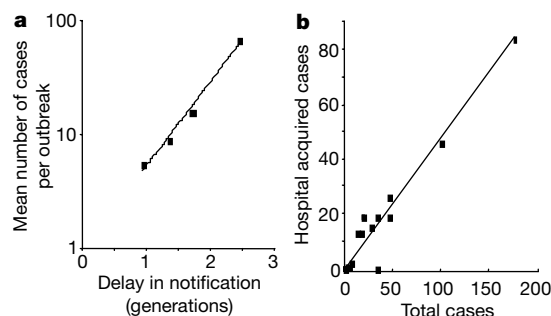


Figure 2 Analysis of 32 European smallpox outbreaks post-1950. **a**, Regression analysis demonstrating the increase in the mean number of cases per outbreak that occurred (note logarithmic scale of ordinate and abscissa) with increased median delay in notification of the outbreak (delay expressed in units of generation time of the disease (19 days)). Regression: $y = 0.95e^{1.71x}$, $r^2 = 0.99$. **b**, Regression analysis showing the relationship between the total number of cases per outbreak and the number of cases that were hospital-acquired. Regression: $y = 0.4777x$, $r^2 = 0.94$.

London¹⁵ or when exacerbated by hospital-associated infections in more recent European outbreaks¹). This analysis is in accordance with inferences from the vaccination level required for smallpox eradication^{1,9} in industrialized countries. The value of 70–80% that was found to be effective would suggest an R_0 between 3.3 and 5 (ref. 9). However, vaccination levels of 80% were not effective in eradicating smallpox in India because of much higher population densities and poorer socio-economic conditions¹⁸, which would imply R_0 values higher than 5. For contemporary outbreaks in industrialized countries, which now have low levels of herd immunity (18%, see Methods), an R_0 value of 4–6 would probably apply for community-acquired infections, but might transiently be higher (for example, 10–12) before the disease was correctly recognized and appropriate hospital infection controls implemented. These values differ from recently applied values. The higher values^{4–6} from 10 to 20 have been estimated from atypical outbreaks with unusually high transmission rates^{1,19}, whereas the lower estimates of around 2 appear to have neglected concurrent herd immunity⁸. Although our estimate for smallpox represents a relatively modest transmission rate by comparison with some other infectious diseases, such as measles or chickenpox⁹, significant epidemics could result, particularly if there were delays in detecting the first cases or in setting up effective public health interventions. □

Methods

Epidemic modelling

For epidemic modelling, transmissibility is often expressed as the basic reproductive number, R_0 , the mean number of secondary cases caused by each primary case in a population composed entirely of susceptible individuals⁹. The reproductive number, R , is an equivalent measure, but without reference to the proportion of susceptible hosts⁹. Smallpox infection results first in a non-infectious incubation period followed by a short prodrome with non-specific symptoms (assumed to be non-infectious²⁰), which together constitute the latency period. Hosts then become overtly infectious and symptomatic, with a pustular rash and, thereafter, slowly recover and are immune, or die^{1,17}. Whilst infectious, hosts transmit disease to susceptible individuals at a rate dependent on R_0 . Public health interventions interrupt transmission by removing infectious individuals into quarantine, or by reducing the susceptibility of contacts by vaccination⁹. These processes were modelled using an appropriately parameterized set of differential equations (1), which were solved using Euler's method with a time-step of 0.1 days.

$$\begin{aligned}
 \frac{dS}{dt} &= \chi_1(1 - \epsilon_1)C_i - \beta(\varphi + \rho - \varphi\rho)SI & \frac{dI}{dt} &= \alpha(1 - \theta)E_n - (\theta + \gamma)I \\
 \frac{dE_n}{dt} &= \beta\varphi(1 - \rho)SI - \alpha E_n & \frac{dQ}{dt} &= \alpha(1 - \epsilon_2)E_i + \theta(\alpha E_n + I) - \chi_2 Q \\
 \frac{dE_i}{dt} &= \beta\varphi\rho SI - (\chi_1\epsilon_2 + \alpha(1 - \epsilon_2))E_i & \frac{dU}{dt} &= \gamma I + \chi_2 Q \\
 \frac{dC_i}{dt} &= \beta\rho(1 - \varphi)SI - \chi_1 C_i & \frac{dV}{dt} &= \chi_1(\epsilon_2 E_i + \epsilon_1 C_i)
 \end{aligned} \tag{1}$$

S and I represent the proportion of the population susceptible and infectious, respectively. U represents the proportion dead and recovered, with the number dead dependent on the specific case fatality rates given in the data sets. The contacts of cases are divided into the following classes: E_u , the number of untraced latent individuals in the population, E_i , the number of traced latent contacts, and C_i , the number of traced uninfected contacts. The final class of contacts are those untraced and uninfected and so effectively remaining in S . Q represents the proportion in quarantine and V the proportion protected by vaccination. The average rate at which latent individuals become infectious^{13,21} is $\alpha = (\text{latency period})^{-1} = 0.0685 \text{ days}^{-1}$ and the rate at which infectious individuals in the community recover or die²² is $\gamma = (\text{infectious period})^{-1} = 0.116 \text{ days}^{-1}$. Two states of quarantine are defined: the first for the traced contacts successfully vaccinated and released into the community at a rate χ_1 , and the second for the infectious cases, which enter U at a rate χ_2 . Different vaccine efficacies are assumed for those uninfected, ϵ_1 , and infected, ϵ_2 . The proportion of contacts found through contact tracing is ρ and the daily rate at which infectious individuals enter quarantine from the community is θ . The proportion of contacts infected is defined as φ . The rate at which potentially infected contacts occur is defined as β , as in equation (2), and N is the size of the population in which the epidemic occurs.

$$\beta = \frac{R_0 \gamma}{\varphi N} \quad (2)$$

Additional assumptions are that no transmission occurs from those quarantined, dead or recovered and the background mortality rate was assumed to be negligible over the time periods examined.

For the Boston, Burford, Warrington and Chester data sets, $\rho = \theta = 0$, which effectively reduces equations (1) above to a simple SEIR model⁹. Intervention parameters were only required when equations (1) was fitted to the data from Kosovo. Here, interventions were implemented 31 days after the onset of symptoms in the index case¹ with the associated parameters shown in Table 2. The number of potentially infected contacts per case was determined as 50 (ref. 1). Values of R_0 were derived for each outbreak by minimizing the mean square error between the mortality data and the predictions of mortality from the model, while applying the outbreak-specific case fatality rates to U and adjusting R_0 and time of onset of symptoms in the index case. In the case of Kosovo, equations (1) were fitted more simply to the reported number of cases rather than deaths. All the other parameters required for equations (1) were obtained independently from the published source(s) given in Table 1. For epidemics in London, R_0 was roughly calculated from the interepidemic interval, $T = 2\pi[L(D + D')/(R_0 - 1)]^{1/2}$, where L is life expectancy between 1840 and 1870 adjusted for excess births over deaths, equal to 25 years, and $D + D'$ is latent + infectious period, equal to 0.063 years².

Estimation of current vaccination coverage

Given that smallpox vaccination ceased in industrialized countries in the mid to late 1970s (ref. 13), a crude estimate of the immunity of the contemporary UK population was calculated, on the basis of 50% having been vaccinated as infants up to 1972, and estimating that about 60% of these would be alive today from current population statistics. Of these only about 60% would still be protected by the vaccinations done on average 50 years previously, calculated by extrapolating from data on secondary attack rates, which increased from 4 to 12% over 10 years following vaccination²³. This suggests that the level of herd immunity may be about 18%, which will continue to decrease with time.

Received 8 October; accepted 5 November 2001.

- Fenner, F., Henderson, D. A., Arita, I., Jezek, Z. & Ladnyi, I. D. *Smallpox and its Eradication* (World Health Organization, Geneva, 1988).
- World Health Organization. *Future Research on Smallpox Virus Recommended* Press Release WHO/77 (World Health Organization, Geneva, 1999).
- US sounds alarm over smallpox weapon threat. *Nature* **399**, 628 (1999).
- Henderson, D. A. *et al.* Smallpox as a biological weapon: medical and public health management. *J. Am. Med. Assoc.* **281**, 2127–2137 (1999).
- Henderson, D. A. Bioterrorism as a public health threat. *Emerg. Infect. Dis.* **4**, 488–492 (1999).
- Henderson, D. A. The looming threat of bioterrorism. *Science* **283**, 1279–1282 (1999).
- O'Toole, T. Smallpox: an attack scenario. *Emerg. Infect. Dis.* **4**, 488–492 (1999).
- Meltzer, M. I., Damon, L., LeDuc, J. W. & Millar, J. D. Modeling the potential responses to smallpox as a bioterrorist weapon. *Emerg. Infect. Dis.* **7** (in the press; also available on <http://www.cdc.gov/ncidod/eid/vol7no6/meltzer.htm>).
- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1992).
- Creighton, C. A *History of Epidemics in Britain* Vol. 2 (Cambridge Univ. Press, Cambridge, 1891).
- Moody, J. The Burford small-pox outbreak of 1758. *Tolsey Pap.* No. 1 (The Tolsey Museum, Burford and Burford School, 1980).
- Registrar General. *Annual Report of the Registrar General 1869–74* (Her Majesty's Stationery Office (HMSO), London, 1870–75).
- Mack, T. M. Smallpox in Europe, 1950–1971. *J. Infect. Dis.* **125**, 161–169 (1972).
- World Health Organization. *Health Aspects of Chemical and Biological Weapons* (WHO, Geneva, 1970).
- Duncan, C. J., Duncan, S. R. & Scott, S. Oscillatory dynamics of smallpox and the impact of vaccination. *J. Theor. Biol.* **183**, 447–454 (1996).
- Dietz, K. & Heesterbeek, J. A. Bernoulli was ahead of modern epidemiology. *Nature* **408**, 513–514 (2000).
- Dixon, C. W. *Smallpox* (Churchill, London, 1962).
- Arita, I., Wickett, J. & Fenner, F. Impact of population density on immunization programmes. *J. Hyg. Cambridge* **96**, 459–466 (1986).

- Wehrle, P. F., Posch, J., Richter, K. H. & Henderson, D. A. An airborne outbreak of smallpox in a German hospital and its significance with respect to other outbreaks in Europe. *Bull. World Health Organ.* **4**, 669–679 (1970).
- Henderson, D. A. & Fenner, F. In *Vaccines* 2nd edn (eds Plotkin, S. A. & Mortimer, E. A.) 13–40 (Saunders, Philadelphia, 1994).
- Gelfand, H. M. & Posch, J. The recent outbreak of smallpox in Meschede, West Germany. *Am. J. Epidemiol.* **94**, 234–237 (1971).
- Koplan, J. P., Azizullah, M. & Foster, S. O. Urban hospitals and rural village smallpox in Bangladesh. *Trop. Geogr. Med.* **30**, 355–358 (1978).
- Mack, T. M., Thomas, D. B., Ali, A. & Khan, M. M. Epidemiology of smallpox in West Pakistan: I. Acquired immunity and distribution of disease. *Am. J. Epidemiol.* **95**, 157–168 (1972).
- Franz, D. R. *et al.* Clinical recognition and management of patients exposed to biological warfare agents. *J. Am. Med. Assoc.* **278**, 399–411 (1997).

Acknowledgements

This work was funded by the Department of Health, UK. The views expressed in the publication are those of the authors and not necessarily those of the Department of Health. We thank C. Penn and G. Lloyd for their help with this work and the preparation of the manuscript, and D. Jones, S. Duncan, N. Gay, and members of the DH Steering Group for their comments and help with model parameterization.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.L. (e-mail: steve.leach@camr.org.uk).

Imperfect vaccines and the evolution of pathogen virulence

Sylvain Gandon^{*†}, Margaret J. Mackinnon^{*†}, Sean Nee^{*} & Andrew F. Read^{*}

^{*} Institute of Cell, Animal and Population Biology, The University of Edinburgh, Edinburgh EH9 3JT, UK

[†] These authors contributed equally to this work

Vaccines rarely provide full protection from disease. Nevertheless, partially effective (imperfect) vaccines may be used to protect both individuals and whole populations^{1–3}. We studied the potential impact of different types of imperfect vaccines on the evolution of pathogen virulence (induced host mortality) and the consequences for public health. Here we show that vaccines designed to reduce pathogen growth rate and/or toxicity diminish selection against virulent pathogens. The subsequent evolution leads to higher levels of intrinsic virulence and hence to more severe disease in unvaccinated individuals. This evolution can erode any population-wide benefits such that overall mortality rates are unaffected, or even increase, with the level of vaccination coverage. In contrast, infection-blocking vaccines induce no such effects, and can even select for lower virulence. These findings have policy implications for the development and use of vaccines that are not expected to provide full immunity, such as candidate vaccines for malaria⁴.

Previous studies on the evolution of vaccine resistance have focused on the spread of 'escape' mutants that display epitopes different to those in the vaccine, thereby escaping immune recognition^{5–7}—this has already happened for polio⁸ and hepatitis B⁹. New vaccines may eventually get around this problem by, for example, targeting conserved epitopes or multiple epitopes simultaneously. Here we study an alternative counter-adaptation to vaccination involving pathogen life-history traits, namely virulence (induced host mortality) and transmission rate. To address this issue we incorporated standard evolutionary theory for virulence evolution^{10–12} into an epidemiological framework¹.

We begin with an analysis of the evolution of parasite virulence

(the disease-induced mortality rate of the host) in a homogeneous host population. We do this by studying the ability of a rare mutant, with virulence denoted α^* , to invade a population of resident parasites with virulence α (the asterisk distinguishes the mutant's trait from the resident's). The evolutionarily stable (ES) pathogen virulence can be found by maximizing the mutant pathogen's $R_0[\alpha^*, \alpha]$ at $\alpha = \alpha^*$. (See equation (1) below.) When the host population is homogeneous and has reached epidemiological equilibrium (as denoted by the circumflex accent ('hat' symbol) throughout), the expression for the mutant's fitness is given by the expected number of secondary cases produced by a single host infected by this mutant over its entire infectious period^{10–13}:

$$R_0[\alpha^*, \alpha] = \frac{\beta^*(\hat{x} + \sigma\hat{y})}{\delta + \alpha^* + \chi^* + \sigma\hat{h}} \quad (1)$$

where x and y are the densities of uninfected and infected hosts, respectively, β is the pathogen's transmission rate, $h = \beta y$ is the rate at which hosts acquire new infections (termed 'the force of infection'), χ is the pathogen's clearance rate (rate at which the host becomes non-infectious), δ is the host's natural mortality rate, and σ is the efficiency with which the pathogen invades an already infected host (superinfection) relative to invading an uninfected host^{13,14}. The superinfection parameter, σ , can also be modelled as a function of virulence¹³, but here, for simplicity, we assume it to be a constant. It is assumed that superinfecting parasites immediately replace the strain already present in the host: thus σh is the rate at which the pathogen is cleared from the host due to arrival of another strain. Note that $\hat{h}$ is determined by the resident pathogen strain. Thus, by setting the density of infected hosts to zero, we recover the classical definition of R_0 , which allows us to tell whether the mutant pathogen can invade a fully susceptible host population¹.

Here we assume, as in classical models of the evolution of virulence^{10–14}, that the pathogen fitness function in equation (1) includes trade-offs involving pathogen virulence—that is, virulence has beneficial, pleiotropic effects on other pathogen life-history traits that offset the fitness cost of host death (which prematurely ends the infectious period). Two types of virulence benefits have been proposed. First, transmission rate is assumed to be an increasing function of pathogen virulence. Second, clearance rate is assumed to be slower with higher virulence. The net result of these negative and positive influences on pathogen fitness is that there is an intermediate optimum level of virulence that maximizes fitness. Although there are few data testing these assumed fitness relationships in pathogens, they are generally supportive^{10,15,16}. The exact nature of these relationships will depend on the biology of each particular host–pathogen interaction, but here we define these trade-offs in simple forms by:

$$\begin{aligned} \beta &= \beta[\alpha] = b_1 \alpha^{b_2} \\ \chi &= \chi[\alpha] = c_1 \alpha^{-c_2} \end{aligned} \quad (2)$$

where the coefficients with subscripts are constants that determine the shape of the trade-offs, and hence the value of α that maximizes fitness.

The question now is how does host immunity (or 'resistance') change the optimum virulence relative to that in a completely non-immune ('susceptible') host population? Still assuming a homogeneous host population, we consider four different forms of immunity, with efficacies denoted r_1 , r_2 , r_3 and r_4 , which independently affect different stages of the pathogen's life cycle (Fig. 1). The first is anti-infection immunity, which decreases the probability that a host becomes infected. The second is anti-growth-rate immunity, which directly reduces virulence and concomitantly affects transmission rate and host recovery. The third is transmission-blocking immunity, which only decreases parasite transmission. The fourth is anti-toxin immunity which directly reduces virulence but, contrary

to anti-growth-rate immunity, does not affect parasite transmission and host recovery rates. This yields:

$$\begin{aligned} \alpha' &= (1 - r_2)(1 - r_4)\alpha \\ \beta' &= (1 - r_3)\beta[(1 - r_2)\alpha] \\ \chi' &= \chi[(1 - r_2)\alpha] \\ h' &= (1 - r_1)\beta'y' \end{aligned} \quad (3)$$

where the prime pertains to immune hosts. Assuming that only the trade-off between virulence and transmission is operating (clearance rate, χ , is a constant), yields the ES virulence:

$$\alpha^* = \frac{b_2(\delta + \chi + \sigma h(1 - r_1)(1 - r_3))}{(1 - b_2)(1 - r_2)(1 - r_4)} \quad (4)$$

Note that, throughout, virulence is measured as induced host mortality in susceptible (non-immune) hosts. Equation (4) implies that anti-growth-rate and anti-toxin immunity (modelled by r_2 and r_4) always select for higher virulence. This is because they reduce the risk of host death and hence selection against more virulent mutants. Indeed, evolution will restore the virulence observed in a uniform population of resistant hosts, as well as the force of infection, to that observed in a uniform population of susceptible hosts by increasing intrinsic virulence (that is, virulence as measured in susceptible hosts). Thus a pathogen following a strategy that would generate optimal virulence in a resistant host will induce a higher-than-optimal virulence in a susceptible host¹⁷. In contrast, anti-infection (r_1) and transmission-blocking (r_3) immunity select for lower virulence whenever there is superinfection, and leave it unchanged otherwise. They act indirectly on the evolution of parasite virulence via the force of infection through their effects on the rate at which an infection is prematurely ended by the arrival of a superinfecting pathogen^{14,17}. Although equation (4) defines the ES virulence as a function of the force of infection, which is itself a function of virulence, the results discussed above can be rigorously proven using implicit differentiation analysis (not shown).

If we alternatively assume that there is only a trade-off between virulence and recovery rate, the ES virulence is:

$$\alpha^* = \frac{(c_1 c_2 / (1 - r_4))^{\frac{1}{1+c_2}}}{1 - r_2} \quad (5)$$

In this case, anti-growth and anti-toxin immunity increase virulence, and the other two forms of immunity have no effect. When both trade-offs are included, the conclusions derived from equation

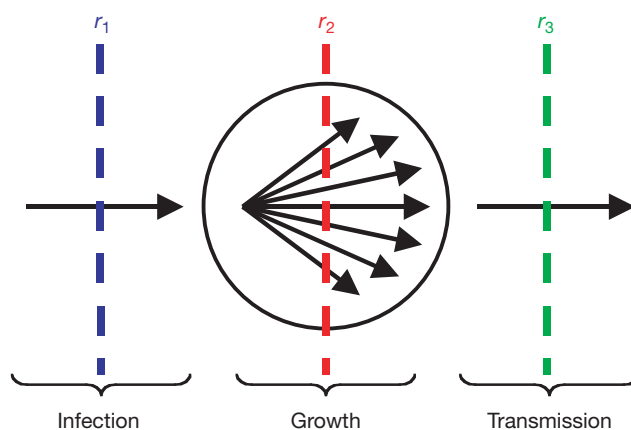


Figure 1 Schematic representation of the action of different types of host resistance at different stages of the pathogen's life cycle. r_1 , anti-infection resistance; r_2 , anti-growth-rate resistance; r_3 , transmission-blocking resistance. A fourth type of resistance—anti-toxin resistance, r_4 —is not shown because it only acts upon host death rates.

(4) do not qualitatively change.

What, then, if the pathogen faces a heterogeneous population of susceptible and resistant hosts, as would happen if a vaccination programme was implemented? On the one hand, we have shown that the consequences of vaccinating individual hosts is to select for higher levels of intrinsic virulence in the case of anti-growth-rate and anti-toxin vaccines. On the other hand, immunized hosts will transmit less, die less and recover more quickly than non-immune hosts, thus reducing the overall level of disease in the population. We now allow this epidemiology to feed back into the pathogen's virulence evolution, and vice versa, in order to determine the overall impact of vaccination programmes on the health of the population. The epidemiological model that we used was a modified version of the standard susceptible-infected model¹ with two classes of hosts—those that are fully susceptible to the pathogen, and those that are partially immune. In addition, we assume a continuous vaccination procedure which provides imperfect but life-long immunity. It is written as:

$$\begin{aligned} dx/dt &= (1-f)\lambda - (\delta+h)x + \chi y \\ dx'/dt &= f\lambda - (\delta+h')x' + \chi'y' \\ dy/dt &= hx - (\delta+\alpha+\chi)y \\ dy'/dt &= h'x' - (\delta+\alpha'+\chi')y' \end{aligned} \quad (6)$$

where λ is a constant rate of flow (which covers both reproduction and immigration) of uninfected hosts into the population, among which a fraction f are resistant, that is, vaccinated, and the forces of infection on susceptible and resistant hosts become $h = \beta y + \beta' y'$ and $h' = (1-r_1)h$, respectively. Note that there are no terms for superinfection in equation (6) as these cancel out. For simplicity, we assume that resistant hosts do not lose immunity to become susceptible, and, except by vaccination, susceptible hosts do not acquire immunity. This latter assumption is relaxed in our malaria example below.

As for the simple homogeneous case, the ES parasite virulence is found by maximizing at $\alpha = \alpha^*$:

$$R_0[\alpha^*, \alpha] = \frac{\beta^*(\hat{x} + \sigma\hat{y})}{\delta + \alpha^* + \chi^* + \sigma\hat{h}} + \frac{\beta'^*(1-r_1)(\hat{x}' + \sigma\hat{y}')}{\delta + \alpha'^* + \chi'^* + \sigma\hat{h}'} \quad (7)$$

which can be seen as a weighted average of the per-host transmission factors¹⁸ on susceptible and resistant hosts (see Supplementary Information). Equations (6) and (7) can then be solved jointly to yield the ES virulence and population prevalence of disease once evolution has occurred.

A numerical example shows that as the efficacy of anti-growth-rate and anti-toxin vaccines increases, there is a marked increase in virulence (Fig. 2a). An exception occurs for very high efficacy anti-growth-rate vaccines, because the contribution to fitness from vaccinated individuals becomes very small (see Supplementary Information). ES virulence always increases with the efficacy of anti-toxin vaccines, because this type of vaccine removes the cost of virulence (increased mortality) without affecting its benefit (increased transmission). Consequently, the fitness contribution of vaccinated hosts always increases with the efficacy of an anti-toxin vaccine. In contrast, as the efficacy of anti-infection and anti-transmission vaccines increases, pathogens will evolve lower levels of virulence if superinfection occurs (Fig. 2a). This is because these vaccines reduce superinfection rates. When pathogens are less likely to be competitively excluded, the benefits of keeping the host alive are greater. The reduction in virulence is weaker for transmission-blocking vaccines than for infection-blocking vaccines because, with transmission-blocking vaccines, vaccinated hosts are fully susceptible to infection. Consequently, the force of infection is higher and this selects for higher virulence levels. When vaccines are fully effective ('perfect'), all except the anti-toxin vaccines share

the same ES virulence. When superinfection occurs, this level falls below the virulence reached when vaccines are never used (or are totally ineffective). This is because, with a perfect vaccine, vaccinated hosts do not transmit the disease and, through the decrease of the force of infection, serve to indirectly favour lower virulence^{17,19}. A perfect anti-toxin vaccine (a vaccine that completely removes the deleterious effects of the parasite on its host) does not decrease transmission and always selects for extreme intrinsic virulence (Fig. 2a).

In addition to virulence consequences, vaccination changes disease prevalence. Figure 2b shows that as vaccination coverage increases, anti-infection and anti-transmission vaccines always reduce disease prevalence when pathogen evolution occurs, and can sometimes eliminate the disease. Anti-growth-rate vaccines, on the other hand, have hardly any effect on prevalence because of a balance between two forces that act in different directions—

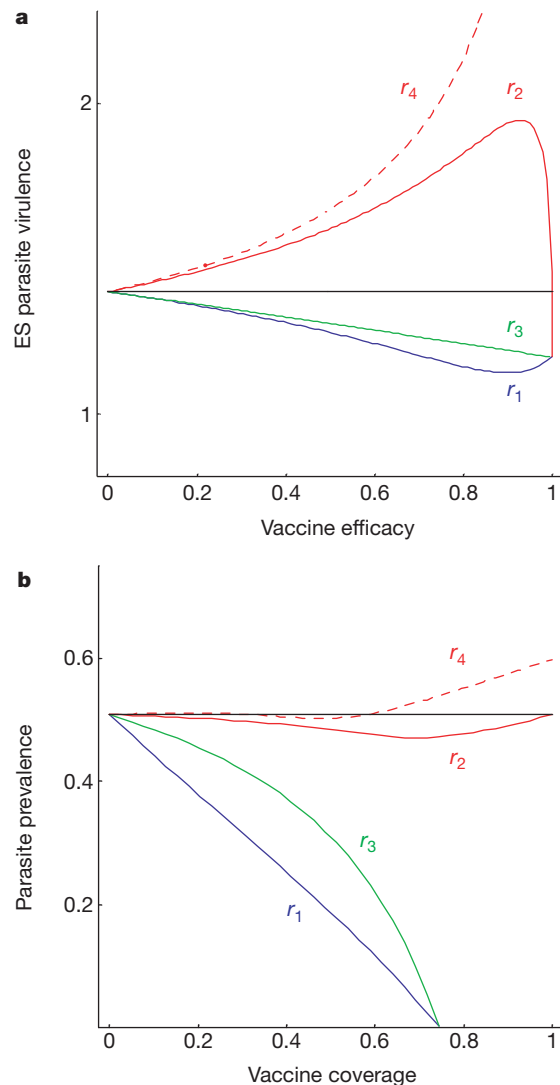


Figure 2 Evolutionary and epidemiological consequences of using different types of vaccines. The different coloured solid curves represent the first three types of vaccines (see Fig. 1), and the red dashed line is for anti-toxin vaccines. **a**, Evolutionarily stable (ES) parasite virulence (measured on susceptible hosts) plotted against the efficacy of the vaccine. **b**, Parasite prevalence (fraction of infected hosts) plotted against the proportion of vaccinated hosts. The horizontal black lines show the outcome in the absence of vaccination. In **a**, $b_1 = 0.5$ and $f = 0.2$ (see text for definitions of b_1 and f). In **b**, $b_1 = 0.2$ and efficacy (r_1 , r_2 , r_3 or r_4) is 0.9. Other parameter values (see text for definitions) are: $\lambda = 25$, $\delta = 1$, $\sigma = 1$, $b_2 = 0.2$, $c_1 = 0$ (that is, no relationship between recovery rate and virulence).

reduced transmission due to the direct effect of the vaccine and increased transmission through the evolution of higher virulence in vaccinated hosts. Anti-toxin vaccines may be even worse. Because these vaccines do not reduce transmission rate, increased vaccination coverage can increase pathogen prevalence above pre-vaccination levels.

The above evolutionary analysis assumes that hosts and pathogens are in population dynamic equilibrium. Extensive numerical simulations indicate that our model generates simple dynamics of rapid approaches to a stable point equilibrium. Many diseases, however, have more complex epidemiological dynamics, which may exhibit cycles or chaotic behaviour. Even if there is a point attractor, the transient dynamics following vaccination may be so pronounced or so long-lasting that an analysis based on equilibrium conditions may be irrelevant. In these situations, the selective pressures would vary both in space (from one host to another) and time, and the evolutionary analysis would need to take into account the effects of such variability on the invasion exponent of a mutant parasite²⁰.

Although a single epidemiological equilibrium exists, the evolutionary analysis of our model revealed that there could be different evolutionary outcomes depending on the initial conditions. Such evolutionary bistability emerges when the parasite can take an evolutionary route that leads to specialization on either susceptible or resistant hosts (leading to low or high virulence, respectively). However, our predictions regarding the effects of imperfect vaccines are not qualitatively altered by evolutionary bistability.

This result has potentially important implications for the evolution of specialization^{21,22} and for the evolution of multihost pathogens²³, but further exploration of this phenomenon falls outside the scope of the present Letter.

Can the above general theory for a virtual pathogen contribute to the rational design of vaccines against real pathogens such as malaria parasites? Current efforts to develop a malaria vaccine are focused on three different stages of the parasite's life cycle—the pre-erythrocytic stages (sporozoites and liver-stage parasites), asexual blood-stage parasites (merozoites and infected erythrocytes) and the mosquito-stage parasites (gametocytes, gametes, ookinetes)⁴. Immunity against these three stages corresponds to the anti-infection, anti-growth-rate and transmission-blocking forms of resistance studied here. Anti-toxin malaria vaccines are also being explored²⁴. Using a modified form of the general model to incorporate two important features of malaria epidemiology—naturally acquired immunity and vector transmission (see Supplementary Information)—we evaluated the public health consequences of using various vaccines.

The model was parameterized using values typical of year-round endemic *Plasmodium falciparum* malaria in a high transmission area. Figure 3a–e shows that, as for the general model, the malaria model predicts that anti-growth-rate and anti-toxin vaccines select for higher virulence, while anti-infection vaccines select for lower parasite virulence. In the malaria model, however, transmission-blocking vaccines may favour slightly higher virulence (Fig. 3c). This vaccine reduces transmission and, consequently, the reproduc-

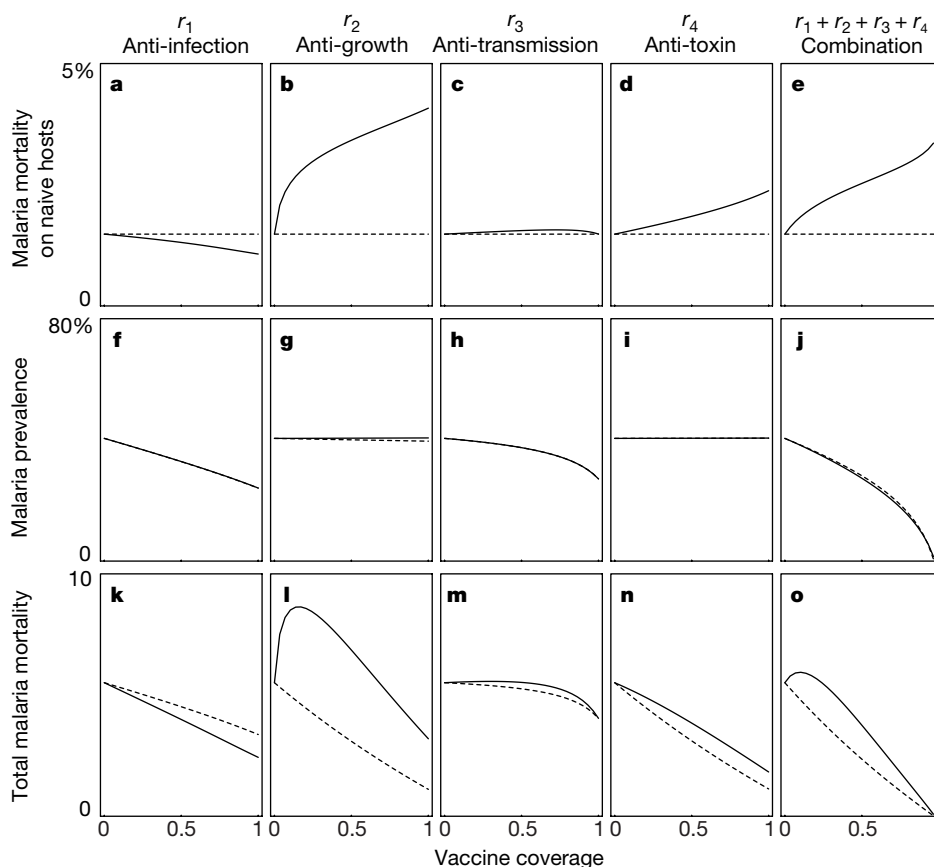


Figure 3 Predicted effects of anti-malaria vaccination coverage on virulence. **a–e**, The probability of dying due to malaria in naive hosts, $\alpha_N/(\alpha_N + \chi_N + \delta)$. **f–j**, the prevalence of malaria ($y_N + y_I + y_V$). **k–o**, The total (population-wide) disease-induced mortality (population average virulence weighted by prevalence of the different host types, given in number of deaths per thousand per year). These results are for an example of endemic malaria; predictions are shown not allowing (dashed line) and allowing (full line) parasite

evolution. We show the effects of single type vaccines and a combination vaccine under the assumption that the different types of vaccines have the same level of efficacy as that provided by natural immunity: $r_1 = \rho_1 = 0.8$, $r_2 = \rho_2 = 0.8$, $r_3 = \rho_3 = 0.8$, $r_4 = \rho_4 = 0.8$. See Supplementary Information for symbol definition and further details on the malaria model.

tive value of parasites infecting vaccinated hosts. In this case, evolution becomes mainly driven by the selective pressures occurring in naturally immune hosts. This explains the increase in virulence in spite of the indirect effect of superinfection acting in the opposite direction (equation (4)). We also examined the evolutionary consequences of a combination of different types of vaccines. The use of a vaccine combining the four different types also favours higher pathogen virulence despite the beneficial effect of anti-infection vaccines (Fig. 3e).

With or without evolution of the pathogen, vaccination is expected to affect the prevalence of malaria. As in the general model (Fig. 2b), the use of anti-infection and transmission-blocking vaccines reduces the force of infection and consequently malaria prevalence (Fig. 3f, h). In contrast, anti-growth-rate and anti-toxin vaccines have hardly any effect on prevalence (Fig. 3g, i). A combination vaccine, via the anti-infection and transmission-blocking effects, is the most efficient in reducing malaria prevalence, and could even lead to eradication for extreme vaccine coverage (Fig. 3j).

The total number of deaths due to malaria depends on both malaria virulence and on the prevalence of infection in the different types of hosts. Figure 3k–o presents the consequences of vaccination at the level of the whole host population. In the absence of pathogen evolution, not surprisingly, all the different types of vaccines decrease the total disease mortality. However, with anti-growth-rate, anti-toxin and transmission-blocking vaccines, evolution towards higher virulence (Fig. 3b–d) erodes the overall benefits of vaccination (Fig. 3l–n). In contrast, when anti-infection vaccines are used, evolution towards lower virulence (Fig. 3a) may increase the population-level benefits of vaccination (Fig. 3k). At high vaccination coverage, a vaccine that incorporates all four types of vaccines would be the most efficient, even when evolution occurs (Fig. 3o). This result supports the development of multivalent, multi-stage vaccines which, it is hoped, will provide greater overall protection than single-target vaccines⁴. Our finding that anti-infection vaccines may have favourable effects on virulence evolution also strongly supports the use of other partially effective control methods (for example, bed nets, mosquito control) to enhance the long-term benefits of vaccination.

How long might such virulence evolution take? Given that genetic variation exists for pathogen virulence^{10,15,16,25,26}, one might expect that, like the evolution of vaccine and drug resistance, the evolution of virulence would occur on timescales that are relevant to public health (decades or less). As an explicit example, we used the malaria model to track the spread of a virulence mutant through time following the start of a vaccination campaign. At 90% vaccine coverage with an anti-growth-rate vaccine of 80% efficacy, it took 38 years for a mutant more than twice as virulent to increase from 1% to 50%, after which it spread towards fixation very rapidly (see Supplementary Information). Numerical simulations, however, indicate that the speed of invasion is very sensitive to the shape of the trade-off function. Accurate predictions of both the invasion dynamics of a virulence mutant and of the evolutionary outcome require further investigation of the shape of these trade-off functions.

Purely epidemiological models have demonstrated that vaccines which are protective for individuals in clinical trials can nonetheless generate unwelcome consequences for a population as a whole^{2,3}. Our incorporation of evolution into the analysis shows that clinically detrimental or beneficial evolution can also occur. The direction of virulence evolution depends critically on the type of vaccine, with several types promoting evolution that increases mortality risk to individual hosts. However, virulence management (an application of darwinian medicine^{27,28}) requires specific models to answer questions regarding specific biological systems. Using malaria as an example, we found that the wide-scale use of even reasonably effective anti-growth-rate and transmission-blocking vaccines

may ultimately do little to relieve community disease levels in malaria-endemic areas. Moreover, the evolution prompted by anti-growth-rate and anti-toxin vaccines can substantially increase the risk for non-immune individuals, such as unvaccinated children and non-immune travellers. The widespread use of such vaccines thus raises difficult ethical issues. Nevertheless, it is probable that anti-disease vaccines (anti-growth-rate and anti-toxin) will be used widely for their short-term beneficial effect at the individual level.

Like drug resistance, the clinically detrimental evolution that we are discussing here will occur on timescales longer than those of clinical trials. Marked increases in virulence of some viral diseases have already followed widespread use of anti-growth-rate vaccines in the chicken industry²⁹. When human populations become uncontrolled experimental systems, we recommend that at the very least, intrinsic virulence of the pathogen population (or more realistically, putative virulence determinants such as *in vitro* multiplication rates³⁰) be closely monitored. □

Received 27 June; accepted 26 October 2001.

- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans* (Oxford Univ. Press, Oxford, 1991).
- McLean, S. A. & Blower, S. M. Imperfect vaccines and herd immunity to HIV. *Proc. R. Soc. Lond. B* **256**, 9–13 (1993).
- Halloran, M. E., Struchiner, C. J. & Spielman, A. Modeling malaria vaccines II: Population effects of stage-specific malaria vaccines dependent on natural boosting. *Math. Biosci.* **96**, 115–149 (1989).
- Hoffman, S. L. *Malaria Vaccine Development: A Multi-immune Response Approach* (American Society for Microbiology Press, Washington DC, 1996).
- McLean, A. R. Vaccination, evolution and changes in the efficacy of vaccines: a theoretical framework. *Proc. R. Soc. Lond. B* **266**, 389–393 (1995).
- Gupta, S., Ferguson, N. M. & Anderson, R. M. Vaccination and the population structure of antigenically diverse pathogens that exchange genetic material. *Proc. R. Soc. Lond. B* **266**, 1435–1443 (1997).
- Lipsitch, M. Vaccination against colonizing bacteria with multiple serotypes. *Proc. Natl Acad. Sci. USA* **96**, 6571–6576 (1997).
- CDC. Public health dispatch: outbreak of poliomyelitis—Dominican Republic and Haiti, 2000. *Morb. Mortal. Wkly Rep.* **46**, 1094 (2000).
- Zuckerman, A. J. Effect of hepatitis B virus mutants on efficacy of vaccination. *Lancet* **356**, 1382–1384 (2000).
- Anderson, R. M. & May, R. M. Co-evolution of hosts and parasites. *Parasitology* **86**, 411–426 (1982).
- Levin, S. A. & Pimentel, D. Selection of intermediate rates of increase in parasite-host systems. *Am. Nat.* **116**, 308–315 (1981).
- Frank, S. A. Models of parasite virulence. *Q. Rev. Biol.* **76**, 37–78 (1996).
- Gandon, S., Jansen, V. A. A. & van Baalen, M. Host life-history and the evolution of parasite virulence. *Evolution* **56**, 1056–1062 (2001).
- Nowak, M. A. & May, R. M. Superinfection and the evolution of parasite virulence. *Proc. R. Soc. Lond. B* **256**, 81–89 (1994).
- Lipsitch, M. & Moxon, E. R. Virulence and transmissibility of pathogens: what is the relationship? *Trends Microbiol.* **6**, 31–36 (1997).
- Mackinnon, M. J. & Read, A. F. Genetic relationships between parasite virulence and transmission in the rodent malaria *Plasmodium chabaudi*. *Evolution* **56**, 689–703 (1999).
- Gandon, S. & Michalakis, Y. Evolution of parasite virulence against qualitative or quantitative host resistance. *Proc. R. Soc. Lond. B* **266**, 985–990 (2000).
- van Baalen, M. & Sabelis, M. W. The dynamics of multiple infection and the evolution of virulence. *Am. Nat.* **146**, 881–910 (1995).
- May, R. M. & Nowak, M. A. Superinfection, metapopulation dynamics, and the evolution of diversity. *J. Theor. Biol.* **176**, 95–114 (1994).
- Caswell, H. *Matrix Population Models: Construction, Analysis, and Interpretation* (Sinauer, Sunderland, 2001).
- Regoes, R. R., Nowak, M. A. & Bonhoeffer, S. Evolution of virulence in a heterogeneous host population. *Evolution* **56**, 64–71 (2000).
- Ronce, O. & Kirkpatrick, M. When sources become sinks: migrational meltdown in heterogeneous habitats. *Evolution* **56**, 1520–1531 (2001).
- Woolhouse, M. E. J., Taylor, L. H. & Haydon, D. T. Population biology of multihost pathogens. *Science* **296**, 1109–1112 (2001).
- Tachado, S. D. et al. Signal transduction in macrophages by glycosylphosphatidylinositols of *Plasmodium*, *Trypanosoma*, and *Leishmania*: activation of protein tyrosine kinases and protein kinase C by inositolglycan and diacylglycerol moieties. *Proc. Natl Acad. Sci. USA* **96**, 4022–4027 (1997).
- Read, A. F. & Taylor, L. H. The ecology of genetically diverse infections. *Science* **296**, 1099–1102 (2001).
- Ebert, D. Experimental evolution of parasites. *Science* **286**, 1432–1435 (1998).
- Williams, G. & Nesse, R. M. The dawn of Darwinian medicine. *Q. Rev. Biol.* **66**, 1–22 (1991).
- Dieckmann, U., Metz, J. A. J., Sabelis, M. W. & Sigmund, K. (eds) *Virulence Management: The Adaptive Dynamics of Pathogen-Host Interactions* (Cambridge Univ. Press, Cambridge, 2001).
- Witter, R. L. Avian tumor viruses: persistent and evolving pathogens. *Acta Vet. Hung.* **46**, 251–266 (1997).
- Chotivanich, K. T. et al. Parasite multiplication potential and the severity of falciparum malaria. *J. Infect. Dis.* **186**, 1206–1209 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank M. van Baalen, R. Carter, D. Ebert, V. Jansen, T. Little, Y. Michalakakis, F. Rousset, and S. West for discussions, and the Leverhulme Trust, BBSRC and the Wellcome Trust for support.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.G. (e-mail: Sylvain.Gandon@ed.ac.uk).

***Drosophila* Toll is activated by Gram-positive bacteria through a circulating peptidoglycan recognition protein**

Tatiana Michel, Jean-Marc Reichhart, Jules A. Hoffmann & Julien Royet

Institut de Biologie Moléculaire et Cellulaire, UPR 9022 du CNRS, 15 rue René Descartes, 67084 Strasbourg cedex, France

Microbial infection activates two distinct intracellular signalling cascades in the immune-responsive fat body of *Drosophila*^{1,2}. Gram-positive bacteria and fungi predominantly induce the Toll signalling pathway, whereas Gram-negative bacteria activate the Imd pathway^{3,4}. Loss-of-function mutants in either pathway reduce the resistance to corresponding infections^{4,5}. Genetic screens have identified a range of genes involved in these intracellular signalling cascades^{6–12}, but how they are activated by microbial infection is largely unknown. Activation of the transmembrane receptor Toll requires a proteolytically cleaved form of an extracellular cytokine-like polypeptide, Spätzle¹³, suggesting that Toll does not itself function as a *bona fide* recognition receptor of microbial patterns. This is in apparent contrast with the mammalian Toll-like receptors¹⁴ and raises the question of which host molecules actually recognize microbial patterns to activate Toll through Spätzle. Here we present a mutation that blocks Toll activation by Gram-positive bacteria and significantly decreases resistance to this type of infection. The mutation *semmelweis* (*seml*) inactivates the gene encoding a peptidoglycan recognition protein (PGRP-SA). Interestingly, *seml* does not affect Toll activation by fungal infection, indicating the existence of a distinct recognition system for fungi to activate the Toll pathway.

To isolate new genes implicated in the *Drosophila* immune response, we screened the first chromosome for mutations that inactivate genes involved in the control of the challenge-induced expression of the antimicrobial peptides Diptericin and Drosomycin. *diphtericin* expression is controlled by the Imd pathway, whereas induction of *drosomycin* is dependent on the Toll pathway^{4,5}. From 2,500 independent lines, we isolated a mutant line in which the ability to express *drosomycin* in response to infection by Gram-positive bacteria (*Micrococcus luteus*, *Streptococcus faecalis*, *Bacillus thuringiensis*) was abolished or severely reduced (Fig. 1a and data not shown). We named this mutation *semmelweis* (*seml*) after Ignaz Philipp Semmelweis, a Hungarian physician who was a pioneer in the field of antiseptic treatments and discovered the cause of puerperal fever¹⁵. Expression of *diphtericin* by Gram-negative bacteria (*Escherichia coli*, *Erwinia carotovora carotovora* and *Enterobacter cloacae*; Fig. 1b and data not shown) was unaffected in *seml* mutant flies but abolished in two mutants of the Imd

pathway^{7,8}, *key* and *dredd* (Fig. 1b). These phenotypes of *seml* in response to challenges by various microorganisms are similar to those of loss-of-function mutants of the Toll pathway^{5,6} (*spz*, *Dif*) but distinct from mutations affecting the Imd pathway.

In addition to its Toll-dependent induction by Gram-positive bacteria, *drosomycin* expression in the fat body is also activated by fungal infection in a Toll-mediated process⁵. When challenging *seml* flies with the entomopathogenic fungus *Beauveria bassiana*, we noted that *drosomycin* expression was wild type, in contrast to the results obtained with *spz* and *Dif* mutants, in which fungal induction of *drosomycin* was nearly abolished (Fig. 1c). *seml* is therefore the first described *Drosophila* mutation that specifically impairs the Toll-dependent induction of *drosomycin* by Gram-positive bacteria without affecting that induced by fungal infection.

We next analysed the resistance of *seml* mutants to infections by various microorganisms. We compared the data with those obtained with *spz* and *key* mutants. The results (Fig. 2) show that *seml* mutants are highly susceptible to Gram-positive infection (*Bacillus megaterium*, *S. faecalis*), but are as resistant as wild-type flies to fungal (*B. bassiana*) and Gram-negative bacterial infection (*E. coli*, *E. c. carotovora*). As expected, in these experiments *key* flies were susceptible only to Gram-negative infection⁸, and *spz* mutants both to fungal and Gram-positive infections⁵.

The similarities between the antibacterial responses in *seml*, *spz* and *Dif* mutants prompted us to study the epistatic relationship between *seml* and the Toll pathway components. We first tested whether *seml* was genetically upstream or downstream of *Toll*, using the *Toll*^{10b} gain-of-function allele, which leads to a challenge-independent expression of *drosomycin*. The levels of *drosomycin* transcription in *Toll*^{10b} and *seml*; *Toll*^{10b} flies were similar (Fig. 3a), indicating that *seml* is genetically upstream of *Toll*. We used the same strategy to analyse the relationship between *seml* and the serine protease inhibitor *nec*. A loss-of-function mutation in the *nec* gene results in challenge-independent expression of *drosomycin* and in the formation of melanotic spots on the cuticle¹³. Both phenotypes are mediated by *spz* and *Toll*. Both *nec* and *seml*; *nec* flies express *drosomycin* in the absence of immune challenge and exhibit melanotic spots (Fig. 3b, c). These results suggest that the *seml* mutation inactivates a protein acting upstream of the Toll receptor and of the protease cascade necessary to activate its ligand Spz through proteolytic cleavage. In contrast to Toll-pathway mutants, homozygous *seml* females are fertile. Therefore *seml* does not seem

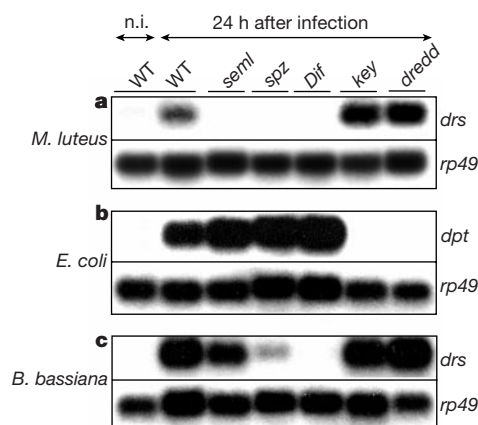


Figure 1 Expression of antimicrobial peptides in different mutant backgrounds after infection by fungi, Gram-positive or Gram-negative bacteria. Northern blots were performed with total RNA from wild type (WT) flies, *seml* mutant flies, or flies mutant for genes in the Toll signalling pathway (*spz*^{tm7}, *Dif*^Δ) and in the Imd pathway (*key*^Δ, *dredd*^{D55}). The flies were infected with *M. luteus* (a), *E. coli* (b) or *B. bassiana* (c) and incubated for 24 h at 25 °C before RNA preparation. *rp49* is used as an RNA loading control. n.i., not induced; *drs*, *drosomycin*; *dpt*, *diphtericin*.

to participate in the Toll-dependent control of dorsoventral axis formation in early embryos¹⁶.

To map the *seml* mutation, we tested the ability of deficiencies in the first chromosome to complement the mutant phenotypes. The *Df(1)HA85* deficiency does not complement *seml* for *drosomycin* inducibility by *M. luteus* (Fig. 4a), nor for survival to Gram-positive infection (Fig. 4b). The phenotypes were identical in homozygous *seml* and hemizygous *seml/Df(1)HA85* females, indicating that *seml* is probably a complete loss-of-function allele. We further observed that *Df(1)N105* complements *seml* (Fig. 4a and data not shown), which narrowed down the chromosomal region containing the mutated gene to the cytological interval 10C01–02; 10F07. To characterize the *seml* mutation, we looked for candidate genes present in this region. Our attention was attracted by the gene coding for a peptidoglycan recognition protein, PGRP-SA, recently shown to bind the peptidoglycan of *M. luteus*^{17–19} used in our screen.

We next compared the sequence of the PGRP-SA complementary DNAs from *seml* and wild-type flies. We noted in the *seml* cDNA the transition of guanine to adenine at position 174, which results in the change of cysteine 80 into a tyrosine (Fig. 4c). This mutation affects a region of amino acids (called the PGRP domain) that is extremely conserved among the members of the PGRP family of vertebrates and invertebrates^{17–19} (Fig. 4c). This cysteine is conserved in more than 90% of the PGRPs, a feature shared by only 5% of the amino

acids that form the PGRP domain. Furthermore, a cysteine is present in close proximity (Cys 74) that might engage in a disulphide bridge with Cys 80. The *seml* mutation would disrupt such a bridge and affect the three-dimensional structure of the protein. To ascertain that the mutation of Cys 80 to Tyr 80 is responsible for the *seml* phenotypes, we undertook rescue experiments by overexpressing wild-type *dPGRP-SA* cDNA, using the ubiquitous driver DaughterlessGal4 (*DaGal4*). Adult *seml; DaGal4-UAS PGRP-SA* flies have a restored ability to induce *drosomycin* after *M. luteus* infection and are no longer susceptible to Gram-positive infection (Fig. 4a, b). In contrast, *seml; DaGal4* flies show phenotypes identical to *seml* flies. These results unambiguously demonstrate that the inability of *seml* mutant flies to resist Gram-positive infection results from the inactivation of the PGRP-SA gene.

The deduced sequence of the PGRP-SA gene indicates the presence of a putative signal peptide¹⁸, suggesting that the PGRP-SA protein is a secreted protein present in the haemolymph or possibly associated with the extracellular matrix. We reasoned that if the PGRP-SA protein is present in the haemolymph, we should be able to rescue the *seml* phenotype by transferring wild-type haemolymph into the mutant flies. Indeed, when wild-type haemolymph was injected into *seml* flies, the recipient flies became capable of expressing *drosomycin* after challenge by *M. luteus* (Fig. 3d). In contrast, *seml* flies injected with *seml* haemolymph did not produce *drosomycin* after immune challenge (Fig. 3e). These results demonstrate that PGRP-SA is a protein circulating in the haemolymph, where the molecular recognition with the Gram-positive bacterial cell wall components can occur.

The data presented above demonstrate that activation of the Toll pathway by Gram-positive bacterial infection is mediated by a circulating peptidoglycan recognition protein. This is the first demonstration of an *in vivo* function for a pattern-recognition receptor in invertebrate innate immunity. Pattern-recognition receptors were defined by Janeway as genome-encoded non-clonally

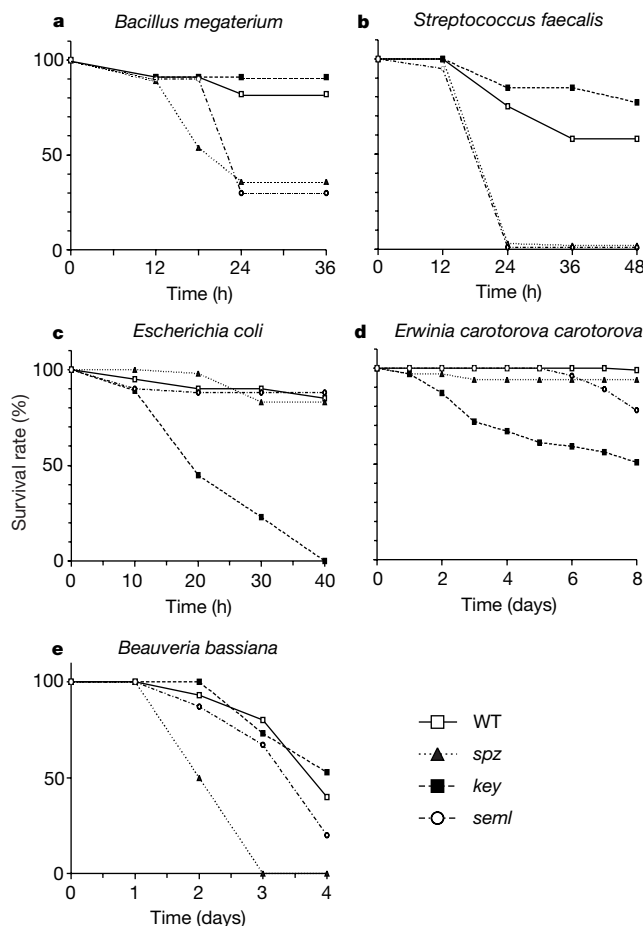


Figure 2 *seml* mutant flies are highly susceptible to Gram-positive infection. The survival rates (%) of wild type (WT), *spz*^{m7}, *key*⁴ and *seml* flies infected with different microorganisms are presented. **a, b**, Gram-positive bacterial infection: *B. megaterium* (**a**) and *S. faecalis* (**b**). **c, d**, Gram-negative bacterial infection: *E. coli* (**c**) and *E. c. carotovora* (**d**). **e**, Fungal infection: *B. bassiana*. Each experiment was performed with 25 flies for each genotype and the results shown are representative of three independent experiments.

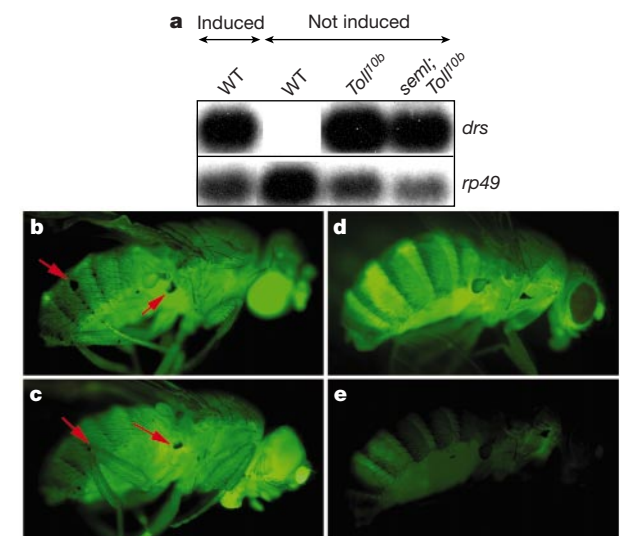


Figure 3 *seml* is genetically upstream of *Toll* and *nec*, and codes for a circulating protein. Genetic epistasis experiments were conducted between *seml*, *Toll* and *nec* mutants. **a**, The constitutive induction of *drosomycin* (*drs*), measured by northern blot) is similar in *Toll*^{10b} and *seml; Toll*^{10b} flies. **b**, *nec*¹/*nec*² flies are characterized by a constitutive expression of *drosomycin* (visualized here with a *drs*-GFP reporter transgene) and by the presence of melanotic spots on the cuticle (arrows). **c**, Both phenotypes are present in the *seml; nec*¹/*nec*² double mutant. **d, e**, *seml; drs*-GFP flies were injected with 20 nl of haemolymph from wild-type flies (**d**) or *seml* mutant flies (**e**). After 1 h, the flies were challenged with *M. luteus*, and 48 h later the induction of *drosomycin* was monitored through GFP expression. The transfer of haemolymph from *seml* flies does not rescue the *seml* phenotype, but haemolymph from wild-type flies does.

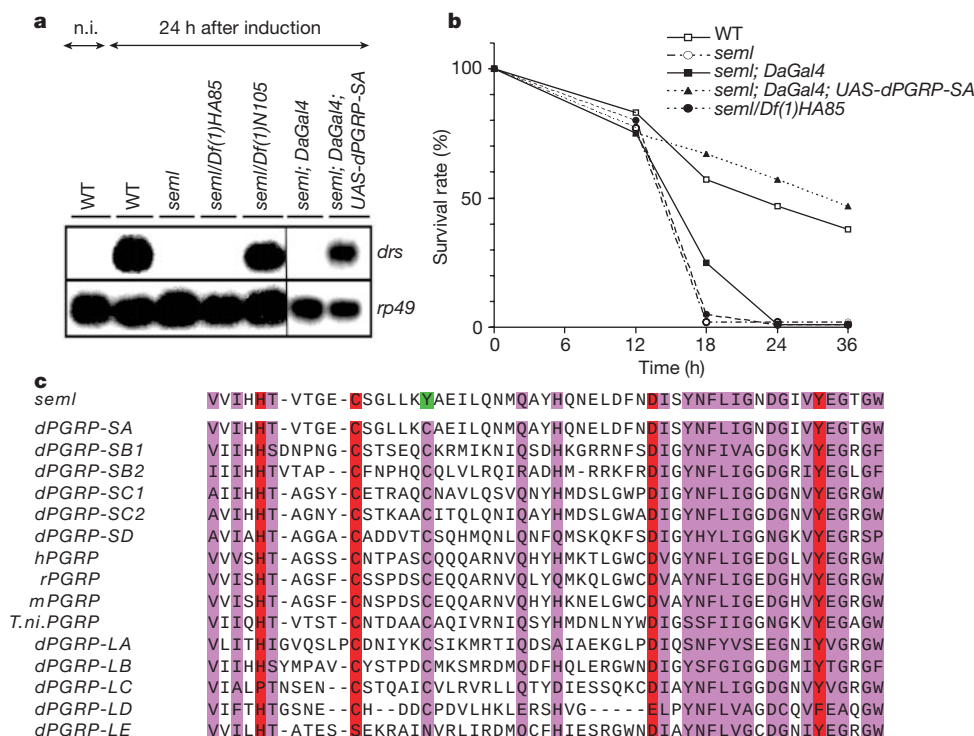


Figure 4 *sem1* is a mutant for the *Drosophila* PGRP-SA gene. **a**, **b**, The *Df(1)HA85* (10C01–02; 11A01–02) deficiency does not complement the *sem1* mutation for *drasomycin* (*drs*) induction after infection by *M. luteus* (**a**) or for the ability to survive infection by *S. faecalis* (**b**). Overexpression of a wild-type (WT) copy of PGRP-SA cDNA is sufficient to rescue the *sem1* phenotype for *drasomycin* induction (**a**) and for survival to *S. faecalis* (**b**). For this purpose the ubiquitous driver *DaGal4* was used. n.i., not induced.

distributed receptors that recognize certain molecular patterns found in microbes but not on self tissues²⁰. The best-documented examples are the various Toll-like receptors present on mammalian immune responsive cells, which bind distinct microbial patterns to activate NF- κ B^{21–26}. Toll-like receptors were discovered in the wake of the demonstration that Toll was involved in the immune response of *Drosophila*²⁷, and it was expected that Toll (and the other receptors of this nine-member family²⁸) would also function as a pattern-recognition receptor in this species. As demonstrated here, this concept applies in the systemic immune response of *Drosophila* not to Toll itself, but to an upstream circulating protein. The existence of proteins susceptible to interact with microbial patterns had been reported from various insect species including *Drosophila*²⁹, but their role in triggering an immune response *in vivo* had not been documented. In view of the relatively large number of genes (12) encoding peptidoglycan recognition proteins in *Drosophila*¹⁸, we were surprised to see that the mutation of one of these genes was sufficient to abolish Toll activation by Gram-positive bacteria. No other circulating protein can substitute for PGRP-SA in the case of *M. luteus* and the three others strains tested. How the recognition of *M. luteus* by PGRP-SA leads to the activation of Toll by cleaving Spätzle remains to be determined in molecular terms. PGRP-SA has no protease domain and we speculate that conformational changes, induced on binding to microbial patterns, may activate associated proteases, analogous to that described for mannose-binding lectin associated proteases (MASPs) in the activation of the lectin pathway³⁰. Finally, the observation that *sem1* mutant flies remain capable of mounting a Toll-dependent antifungal response clearly shows that the upstream mechanisms leading to activation of Toll are specific for each class of microorganisms. The versatility of the Toll-like receptors on mammalian immune-responsive cells is probably

paralleled by a versatility of circulating pattern-recognition receptors in *Drosophila*.

Methods

Fly strains

Stocks were raised on standard cornmeal–agar medium at 25°C. *Oregon R* and *y w drs-GFP dpt-LacZ* flies were used as controls. *Spz^{mt}*, *Dif²*, *key⁴*, *dredd²⁵⁵*, *nec¹* and *Toll^{10b}* alleles are described elsewhere^{6–8,13}. *Df(1)HA85/FM7c*, *Df(1)N105/FM6* and *DaGal4* stocks were obtained from the Bloomington stock centre.

Mutagenesis

y w drs-GFP dpt-LacZ male flies were treated with ethylmethane sulphonate (EMS) and crossed with *C(1)DX y w* females. Single F₁ males were crossed to FM7 balancer-carrying females to establish independent lines. Six females from each line were challenged with a needle dipped into a mix (1/1, vol/vol) of concentrated cultures (absorbance $A_{600} = 200$) of *M. luteus* (CIP:A270) and *E. coli* (1106). The flies were maintained for 48 h at 25°C before induction of GFP and LacZ was monitored by a fluorometric assay in 96-well microplates. *Dif²* and *key⁴* were respectively used as controls for *drasomycin* and *dptericin* induction.

Septic injury and survival experiments

Bacterial infections were performed by challenging adult flies with a thin tungsten needle previously dipped into a concentrated culture of the appropriate bacterial strains. Natural bacterial infections were conducted by incubating adult flies for 12 h in a vial containing a cotton wad soaked with a mix (1/1, vol/vol) of yeast paste and a concentrated solution of *E. c. carotorova*³. To test antifungal response, anaesthetized flies were manually shaken for 30 s on a Petri dish containing a sporulating *B. bassiana* culture. Infected animals were incubated at 25°C, except for fungal infections, where flies were maintained at 29°C. Survival experiments were carried out with 25 flies for each tested genotype. Surviving flies were transferred daily into fresh vials and counted. Results are expressed as percentage of infected flies at different time points after infection. Each experiment is representative of at least three independent experiments. For northern blot analysis, flies were collected 24 h after infection and stored at –80°C.

RNA preparation and northern blot

Total RNA extraction and northern blot experiments were performed as described⁸.

Sequencing of the PGRP-SA allele

DNA from single homozygous flies was extracted using a 'single fly preparation protocol'. The PGRP-SA open reading frame from *seml* and *y w drs-GFP dpt-LacZ* flies was obtained by PCR using Platinum high-fidelity polymerase (Life Technology) and the following primers: 5'-GATAAATCCGGCAGATAGCCCA-3' and 5'-ACTTTACTAACTGATATGCTC-3'. Single PCR reactions were analysed on 2% agarose gel, mixed together and purified with QIAquick PCR purification kit (Qiagen). DNA was sequenced by Eurogentec (Seraing).

Cloning and transformation of *UAS-dPGRP-SA*

The wild-type *PGRP-SA* cDNA was obtained by PCR using *y w drs-GFP dpt-LacZ* cDNA as a template. *EcoRI* and *XbaI* sites were introduced 5' and 3' respectively of the *PGRP* cDNA using the following primers: 5'-GGAATTCCATGCAGCCGGTTCGATTC-3' and 5'-GCTCTAGAGCTTAGGGATTGAGAGCCA-3' and the Platinum high-fidelity polymerase. This fragment was then subcloned into *pUAST*. After sequencing, the construct was injected into *w¹¹¹⁸* embryos.

Received 10 September; accepted 5 October 2001.

- Hoffmann, J. A., Kafatos, F. C., Janeway, C. A. & Ezekowitz, R. A. Phylogenetic perspectives in innate immunity. *Science* **284**, 1313–1318 (1999).
- Khush, R. S., Leulier, F. & Lemaitre, B. *Drosophila* immunity: two paths to NF- κ B. *Trends Immunol.* **22**, 260–264 (2001).
- Lemaitre, B., Reichhart, J. M. & Hoffmann, J. A. *Drosophila* host defense: differential induction of antimicrobial peptide genes after infection by various classes of microorganisms. *Proc. Natl Acad. Sci. USA* **94**, 14614–14619 (1997).
- Lemaitre, B. *et al.* A recessive mutation, *immune deficiency (imd)*, defines two distinct control pathways in the *Drosophila* host defense. *Proc. Natl Acad. Sci. USA* **92**, 9465–9469 (1995).
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M. & Hoffmann, J. A. The dorsoventral regulatory gene cassette *spatzle/Toll/cactus* controls the potent antifungal response in *Drosophila* adults. *Cell* **86**, 973–983 (1996).
- Rutschmann, S. *et al.* The Rel protein DIF mediates the antifungal but not the antibacterial host defense in *Drosophila*. *Immunity* **12**, 569–580 (2000).
- Leulier, F., Rodriguez, A., Khush, R. S., Abrams, J. M. & Lemaitre, B. The *Drosophila* caspase Dredd is required to resist Gram-negative bacterial infection. *EMBO Rep.* **1**, 353–358 (2000).
- Rutschmann, S. *et al.* Role of *Drosophila* IKK γ in a Toll-independent antibacterial immune response. *Nature Immunol.* **1**, 342–347 (2000).
- Lu, Y., Wu, L. P. & Anderson, K. V. The antibacterial arm of the *Drosophila* innate immune response requires an I κ B kinase. *Genes Dev.* **15**, 104–110 (2001).
- Vidal, S. *et al.* Mutations in the *Drosophila* *dTAK1* gene reveal a conserved function for MAPKKKs in the control of rel/NF- κ B-dependent innate immune responses. *Genes Dev.* **15**, 1900–1912 (2001).
- Meng, X., Khanuja, B. S. & Ip, Y. T. Toll receptor-mediated *Drosophila* immune response requires Dif, an NF- κ B factor. *Genes Dev.* **13**, 792–797 (1999).
- Hedengren, M. *et al.* Relish, a central factor in the control of humoral but not cellular immunity in *Drosophila*. *Mol. Cell* **4**, 827–837 (1999).
- Levashina, E. A. *et al.* Constitutive activation of Toll-mediated antifungal defense in serpin-deficient *Drosophila*. *Science* **285**, 1917–1919 (1999).
- Akira, S., Takeda, K. & Kaisho, T. Toll-like receptors: critical proteins linking innate and acquired immunity. *Nature Immunol.* **2**, 675–680 (2001).
- Celine, L. F. *Semmelweis. L'imaginaire* (Gallimard, Paris, 1999).
- Belvin, M. P. & Anderson, K. V. A conserved signaling pathway: the *Drosophila* Toll-dorsal pathway. *Annu. Rev. Cell Dev. Biol.* **12**, 393–416 (1996).
- Kang, D., Liu, G., Lundström, A., Gelius, E. & Steiner, H. A peptidoglycan recognition protein in innate immunity conserved from insects to humans. *Proc. Natl Acad. Sci. USA* **95**, 10078–10082 (1998).
- Werner, T. *et al.* A family of peptidoglycan recognition proteins in the fruit fly *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **97**, 13772–13777 (2000).
- Liu, C., Gelius, E., Liu, G., Steiner, H. & Dziarski, R. Mammalian peptidoglycan recognition protein binds peptidoglycan with high affinity, is expressed in neutrophils, and inhibits bacterial growth. *J. Biol. Chem.* **275**, 24490–24499 (2000).
- Janeway, C. A. Jr Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).
- Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. *Science* **282**, 2085–2088 (1998).
- Poltorak, A., Ricciardi-Castagnoli, P., Citterio, S. & Beutler, B. Physical contact between lipopolysaccharide and Toll-like receptor 4 revealed by genetic complementation. *Proc. Natl Acad. Sci. USA* **97**, 2163–2167 (2000).
- Takeuchi, O. *et al.* Differential roles of TLR2 and TLR4 in recognition of Gram-negative and Gram-positive bacterial cell wall components. *Immunity* **11**, 443–451 (1999).
- Hemmi, H. *et al.* A Toll-like receptor recognizes bacterial DNA. *Nature* **408**, 740–745 (2000).
- Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
- Ozinsky, A. *et al.* The repertoire for pattern recognition of pathogens by the innate immune system is defined by cooperation between Toll-like receptors. *Proc. Natl Acad. Sci. USA* **97**, 13766–13771 (2000).
- Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A. Jr A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. *Nature* **388**, 394–397 (1997).
- Tauszig, S., Jouanguy, E., Hoffmann, J. A. & Imler, J. L. Toll-related receptors and the control of antimicrobial peptide expression in *Drosophila*. *Proc. Natl Acad. Sci. USA* **97**, 10520–10525 (2000).
- Franz, N. C. & White, K. Innate recognition systems in insect immunity and development: new approaches in *Drosophila*. *Microbes Infect.* **2**, 243–250 (2000).
- Turner, M. W. & Hammas, R. M. Mannose-binding lectin: structure, function, genetics and diseases associations. *Rev. Immunogenet.* **2**, 305–322 (2000).

Acknowledgements

We thank our colleagues for comments on the manuscript, M. E. Moritz and M. Schneider

for providing bacterial and fungal cultures, and C. Chevalier for sequencing. This work was supported by CNRS, the Ministère de l'Education Nationale de la Recherche et de la Technologie, and the Fondation pour la Recherche Médicale (Implantation jeunes équipes to J.R.). Financial support from Entomed, Exelixis and the National Institutes of Health is acknowledged.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R. (e-mail: J.Royet@ibmc.u-strasbg.fr).

Apol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*

Miguel Navarro* & Keith Gull

School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Oxford Road, Manchester M13 9PT, UK

In the mammalian host, African trypanosomes generate consecutive waves of parasitaemia by changing their antigenic coat. Because this coat consists of a single type of variant surface glycoprotein (VSG), the question arises of how a trypanosome accomplishes the transcription of only one of a multi-allelic family of VSG expression site loci to display a single VSG type on the surface at any one time¹. No major differences have been detected between the single active expression site and the cohort of inactive expression sites². Here we identify an extranucleolar body containing RNA polymerase I (pol I) that is transcriptionally active and present only in the bloodstream form of the parasite. Visualization of the active expression site locus by tagging with green fluorescent protein³ shows that it is specifically located at this unique pol I transcriptional factory. The presence of this transcriptional body in postmitotic nuclei and its stability in the nucleus after DNA digestion provide evidence for a coherent structure. We propose that the recruitment of a single expression site and the concomitant exclusion of inactive loci from a discrete transcriptional body define the mechanism responsible for VSG mono-allelic expression.

African trypanosomes are responsible for sleeping sickness in humans, an epidemic disease currently affecting up to half a million people. The parasite *Trypanosoma brucei* alternates between a mammalian host and the tsetse vector during its life cycle. The bloodstream form can undergo antigenic variation by switching VSG surface coats¹, thus avoiding the host immune response and ensuring a persistent infection. To achieve the expression of a single type of VSG on the surface, only 1 out of 20 possible telomeric VSG expression sites (ESs) is expressed at a given time². A set of ES-associated genes⁴, which confers selective advantages in particular hosts^{4,5}, are transcribed polycistronically together with the VSG from a promoter located 40–60 kilobases (kb) upstream^{6,7}. Hence, VSG expression site regulation is a genuine example of a mono-allelic expression from a multi-allelic gene family in an early branched eukaryote⁸.

In addition to DNA recombination involving individual VSG genes⁹, antigenic variation can occur by a coupled transcriptional activation/inactivation of ESs, named *in situ* switching. Once active, the choice of this ES is maintained over many generations. Previous

* Present address: Instituto de Parasitología y Biomedicina 'López-Neyra', CSIC, C/ Ventanilla 11, 18001 Granada, Spain.

work has shown that ES *in situ* switching can occur epigenetically². In the procyclic form (the tsetse mid-gut stage of the parasite), no VSG is expressed and an invariant glycoprotein, procyclin, covers the parasite surface¹⁰. Notably, RNA polymerase I (pol I) transcribes not only the ribosomal DNA but also the genes of these two main surface proteins, VSG and *procyclin*, which are characteristic of the different developmental stages^{11–14}.

Although much molecular biology has been done to identify differences between active and inactive ESs in the bloodstream form, no principal differences have been found². Because ES transcription is accomplished by a pol I activity that is normally associated with rDNA expression in the nucleolus¹⁵, we investigated whether VSG ES regulation is mediated by a nuclear compartmentalization process.

We developed antibodies reactive to the large subunit of *T. brucei* pol I. Both rabbit polyclonal (affinity-purified) and mouse monoclonal antibodies recognized a polypeptide in protein extracts from the two developmental stages that is similar to the predicted relative molecular mass (M_r) of the pol I large subunit (196,000) (Fig. 1a). Immunofluorescence analysis using these antibodies in procyclic cells showed that pol I localized to the nucleolus, as expected for a eukaryote¹⁵ (Fig. 1b). But immunofluorescence analysis of the bloodstream form identified a pol-I-containing body in addition to the nucleolus (in 59% and 62% of G1 cells ($n = 500$) using mono- and polyclonal antibodies, respectively) (Fig. 1c). Throughout this work, we assume that this body will not be seen in cells where it lies

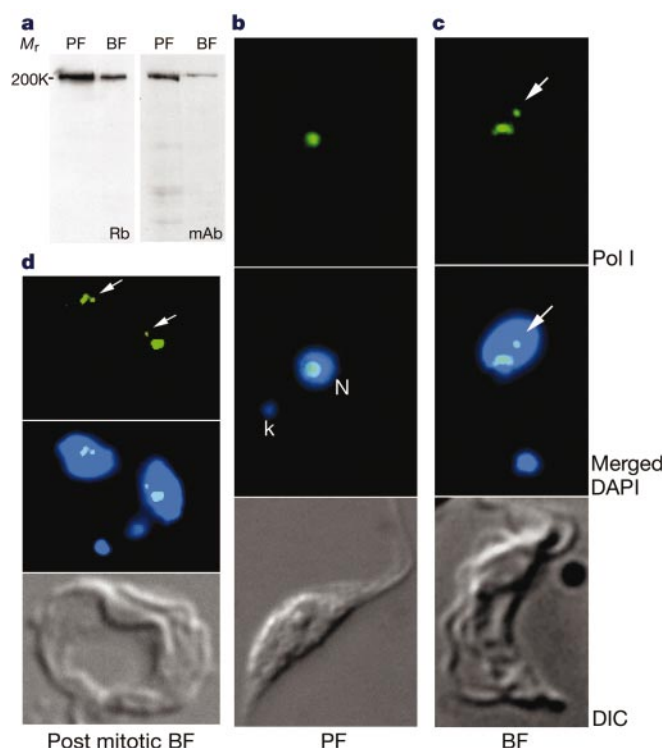


Figure 1 Identification of an extranucleolar body containing pol I in bloodstream-form nuclei of *T. brucei*. **a**, Western blot analysis using rabbit polyclonal (affinity-purified) (Rb) and mouse monoclonal 16B1a (mAb) antibodies reactive to the pol I large subunit. Total protein extracts from procyclic (PF) and bloodstream forms (BF) (5×10^6 cells) were used. **b**, Immunofluorescence analysis of the procyclic form, a tsetse stage of the parasite, using the pol I monoclonal antibody (green; top), merged with DNA stained by DAPI (blue; middle), and differential interference contrast (DIC) image of the cell (bottom). Nuclear DNA (N) and mitochondrial DNA (k) are labelled in a G1 cell. **c**, Immunofluorescence analysis of the wild-type bloodstream form using the pol I monoclonal antibody (green), merged with DNA stained by DAPI (blue), and the DIC image of the cell. Arrows indicate the extranucleolar pol-I-containing body. **d**, Immunofluorescence analysis of a bloodstream-form cell in postmitotic cells where both nuclei are visible and a pol-I-containing body in each nucleus is present (arrows).

above, below or very close to the nucleolus. A detailed reconstruction of an individual nucleus using the Deltavision system shows clearly the discrete three-dimensional nature of the pol I body in the nucleoplasm (see Supplementary Information).

To investigate whether this body represented an additional site of ribosomal RNA synthesis in the nucleus of the bloodstream form, we performed RNA fluorescence *in situ* hybridization (FISH) using the rDNA locus as a probe (Fig. 2a). RNA FISH experiments showed only one area of rRNA synthesis in both procyclic and bloodstream-form nuclei that corresponded to the nucleolus. We also analysed the localization of the well-characterized nucleolar marker fibrillarin, which is involved in one of the pre-splicing reactions of 18S rRNA¹⁶. Fibrillarin was localized to the nucleolus in both procyclic and bloodstream forms, but it did not localize to the pol I extranucleolar body of bloodstream-form nuclei (Fig. 2b, c). Thus, this extranucleolar pol-I-containing body represents a specialized nuclear compartment with a function different to that of the nucleolus.

To determine whether this pol I extranucleolar body was transcriptionally active, we labelled nascent RNA with 5-bromo-UTP (Br-UTP) in permeabilized cell nuclei¹⁷. Many transcriptional foci were detected in a bloodstream-form trypanosome nucleus (Fig. 3a); however, when α -amanitin was included in the buffer to inhibit transcription by pol II and pol III, the foci were reduced to a maximum of two (Fig. 3b). Double-labelling experiments of nascent Br-UTP RNA and pol I confirmed that both of the α -amanitin-resistant transcriptional foci colocalized with pol-I-containing structures (Fig. 3c). These data indicate that in addition to the expected transcription activity of the nucleolus, this unique pol I body is transcriptionally active. Run-on assays in *T. brucei*

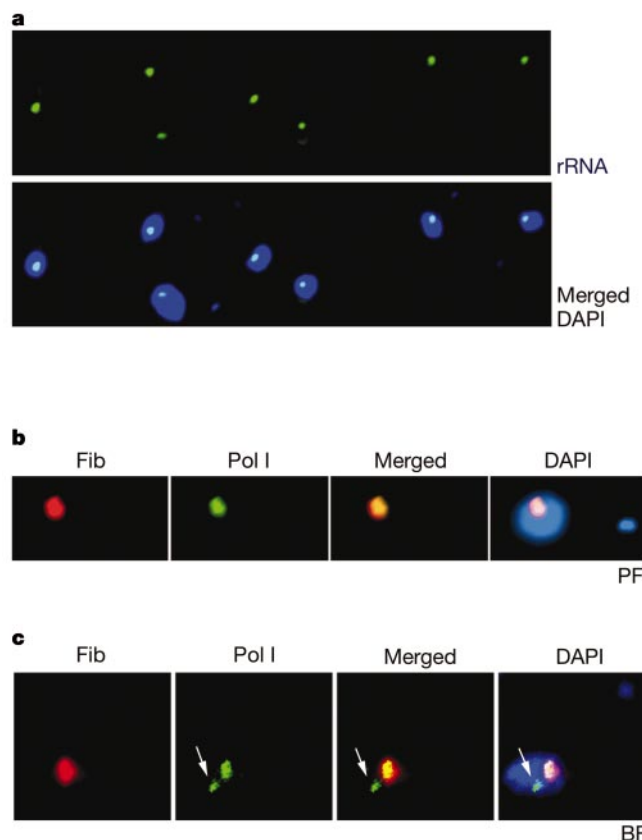


Figure 2 The pol I extranucleolar body has a different function from that of the nucleolus. **a**, Fluorescence *in situ* hybridization using a rDNA as a probe (green; top), and merged with nuclear DNA stained with DAPI, showing a unique signal in each nucleus (bottom). **b**, Double-labelling experiment using an anti-fibrillarin monoclonal antibody (P2G3) and the rabbit polyclonal anti-pol I on a procyclic-form (PF) cell. **c**, Double-labelling experiment as in **b**, but using bloodstream-form (BF) cells, which shows that the pol I extranucleolar body does not contain fibrillarin (arrows).

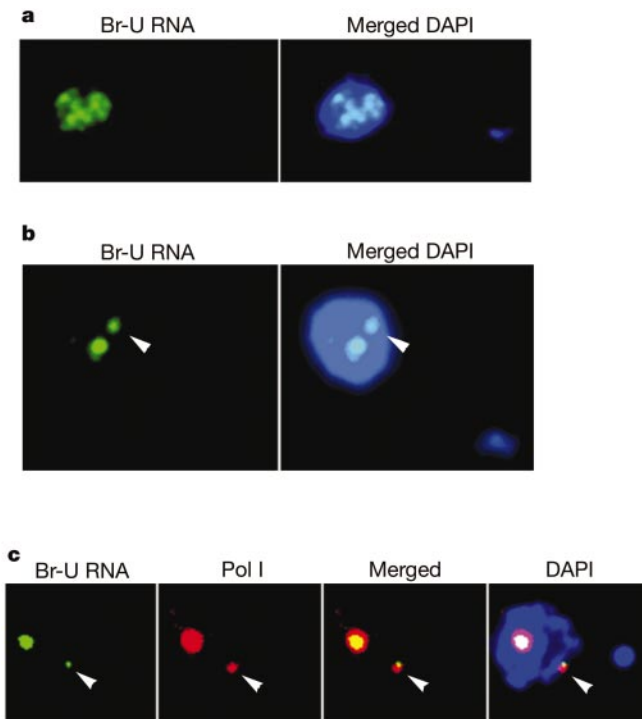


Figure 3 Br-UTP labelling of nascent RNA in permeabilized bloodstream-form nuclei indicates that the extranucleolar pol I structure is transcriptionally active. **a**, Total transcription, as revealed by Br-UTP (Br-U) incorporation foci (green) using a monoclonal antibody reactive to BrdUTP. **b**, Two transcriptional foci detected in the presence of α -amanitin ($100 \mu\text{g ml}^{-1}$). **c**, Colocalization of the α -amanitin-resistant nascent RNA Br-UTP-labelled foci (green) with the nucleolar and extranucleolar body (arrowhead) containing pol I (red).

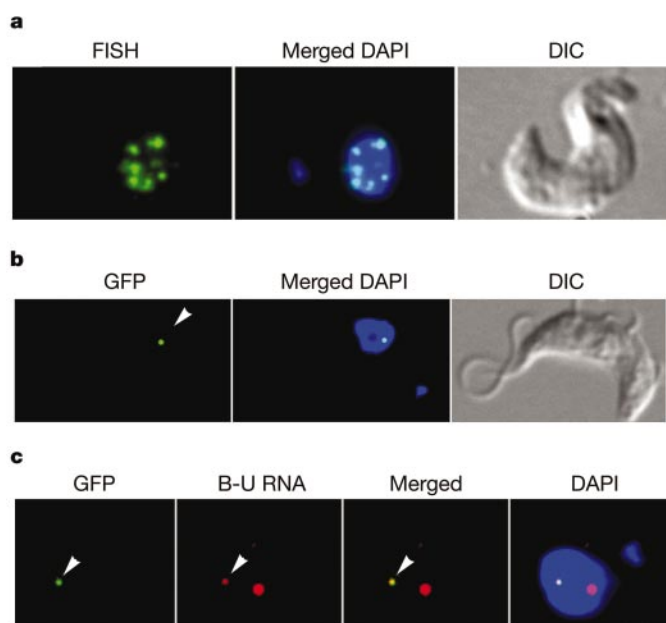


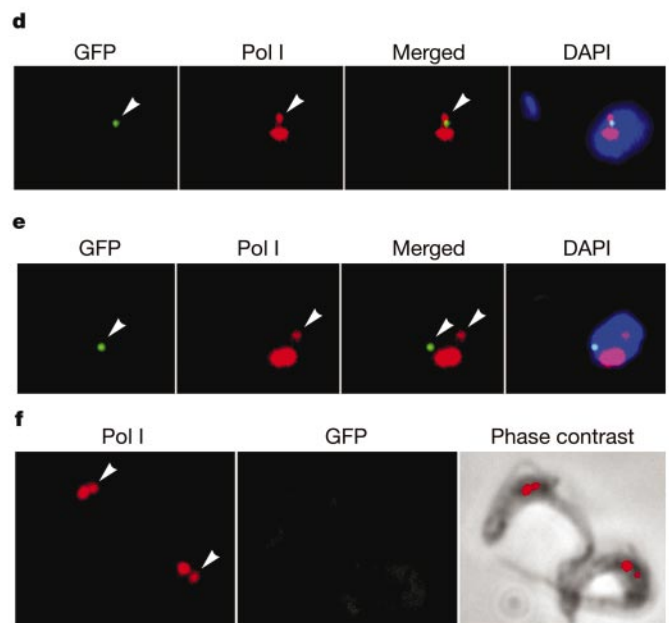
Figure 4 GFP-LacI tagging of the active expression site (ES) sequences in bloodstream-form nuclei reveals the association of this locus with α -amanitin-resistant transcription and with the extranucleolar pol I body. **a**, Distribution of ESs in the nucleus of a bloodstream form detected by DNA FISH using a probe containing the 50-bp repeat sequences that lie upstream of all ESs. **b**, GFP-LacI bound to the ES sequences (green) in a G1 cell detected using an anti-GFP polyclonal antibody. **c**, The small α -amanitin-resistant transcriptional foci (Br-U RNA; red) colocalized with the GFP-tagged active ES sequences (GFP; green). **d**, The extranucleolar pol I body is associated with the GFP-

have indicated that the rDNA and the VSG ESs are transcribed by an α -amanitin-resistant polymerase^{11,12,18} that shares biochemical characteristics with pol I (refs 13, 14). We provide here direct evidence for pol-I-mediated α -amanitin-resistant transcription *in vivo*.

DNA FISH using a probe common to all ESs showed their dispersed distribution in the nucleus of the bloodstream form (Fig. 4a), similar to previous results¹⁹. However, we wanted to determine whether the active VSG ES sequences were associated with the pol I transcriptional body. We therefore established another way to visualize the position of DNA sequences in the nucleus that has been successful in other organisms³. The active 221ES site was tagged with the *lac* operator sequences, thus allowing specific recognition and visualization through binding of the *lac* repressor (LacI) fused to green fluorescent protein (GFP)³. We generated a construct to target the region upstream of the active ES promoter with 256 tandem copies of the operator for the *Escherichia coli* lactose operon. The GFP-LacI fusion protein was expressed in a tetracycline-inducible system²⁰. On tetracycline induction of the transgenic trypanosomes, we detected a single dot in G1 nuclei using anti-GFP antibodies (Fig. 4b).

We examined whether the tagged ES was associated with the pol I transcriptional body. First, double-labelling experiments using anti-GFP antibodies and Br-UTP RNA labelling showed colocalization of the α -amanitin-resistant transcripts and the tagged active ES (Fig. 4c). Second, the GFP-tagged active ES was associated with the pol-I-containing extranucleolar body (Fig. 4d). As a control, we analysed the nuclear position of an inactive ES by inserting the *lac* operator cassette upstream of a silent ES (121ES); this construct showed a lack of association with the pol I transcriptional body (Fig. 4e). The restricted localization of α -amanitin-resistant nascent transcripts, pol I and the active ES (but not an inactive ES) is suggestive of a higher order nuclear architecture.

We examined whether our observations were more than the mere



tagged active ES sequences, detected by double-labelling using monoclonal anti-pol-I (red) and polyclonal anti-GFP (green) antibodies (arrowheads), in the 221ES Lac-operator-tagged cell line. **e**, The extranucleolar pol I body is not associated with the GFP-tagged inactive ES sequences in the 121ES Lac-operator-tagged cell line (arrowheads). **f**, The extranucleolar pol I body is present in the absence of DNA (arrowheads). In contrast, the GFP-LacI signal associated with the active ES was not present after DNase I treatment. Complete digestion of DNA was confirmed by lack of DAPI staining (see Methods).

consequence of resident pol I on the active ES DNA. We compared the stability of the pol I body after DNase I digestion of the nuclei with that of the GFP–LacI bound to the 250 operator repeats upstream of the active ES. This GFP–LacI immunofluorescence dot was completely removed by DNase I treatment, as compared with untreated cells (200 counted; see Methods). However, the pol I body was still detected as a pol-I-positive nucleolar structure (Fig. 4f). This result indicates that there is a higher order architecture, and we propose that this extranucleolar structure should be termed the ‘ES body’ (ESB).

The ESB was solely associated with the expression of the VSG ES in the bloodstream form, a finding given more significance by its absence in the procyclic-tsetse form in which *procyclin* genes in addition to rDNA are also transcribed by pol I. The ESB is only present in the bloodstream form, in which the mutually exclusive expression of a single ES is crucial to display only one VSG type on the surface at any one time. This presence indicates that the ESB structure may have a specific role in the specialized regulation of VSG ESs. In contrast to the GFP–LacI dot, the ESB pol I molecules at the active ES remained after DNase I treatment of the nuclei (Fig. 4f). The presence of the ESB only in the bloodstream form and even in the absence of DNA strongly supports the existence of a coherent architectural body.

The identification of the ESB explains the unexpected localization of nuclear ES transcripts outside the nucleolus¹⁹. The detection of adjacently located nuclear transcripts from two different ESs in a switching process²¹ can also be explained by the existence of a single structure (the ESB) with which each ES alternatively interacts. The transcription and splicing of messenger RNAs are coupled processes. It seems likely that the ESB will contain the RNA processing machinery, providing an explanation for efficient processing of transcripts from the active ES²². We have also detected the ESB in postmitotic nuclei at a point when the DNA has segregated into two nuclei but cytokinesis is not yet complete (Fig. 1d). These data suggest that the ESB is replicated along with the active ES, and that each daughter inherits an ESB with an attached active ES. This process would provide a mechanism for the maintenance and inheritance of the active transcriptional state.

During antigenic switching, the currently accepted ES cross-talking hypothesis² predicts mutual exclusion, with activation of a new ES dependent on inactivation of the currently expressed site, resulting in a coupled process²³. Experiments that attempt to force the expression of two ESs, each targeted with a selectable marker, consistently show that two ESs cannot be fully active at any one time²¹. There is little evidence for chromatin-silencing mechanisms that can repress all inactive ES telomeres except one²; instead, inactive ESs are transcriptionally accessible²³. Given previous data and our current results, we propose a model in which ESB-dependent ES recruitment leads to the activation of a single ES, and inactive ESs are excluded from this structure owing to the presence of the active one. Currently we do not know the mechanism that ensures that only one ES occupies the ESB; some conserved ES sequences may function as a *cis*-acting unique attachment site, similar to the way in which locus control regions²⁴ may function. Our model implies further that *in situ* switching is mediated by an infrequent process whereby the attachment of the active ES with the ESB becomes unstable, thus allowing occupancy by an inactive ES that displaces the active one, resulting in a coupled activation/inactivation.

It has been suggested that multiple transcriptional complexes can be organized into macro-structures, termed transcriptional factories²⁵. But the identification of individual factories is extremely difficult owing to the complexity of pol II and III transcription in eukaryotic nuclei. The ESB that we have identified represents direct evidence for the existence of a highly defined individual transcription factory. Our model, which links such factories to the expression of particular genes, may help to explain

other mono-allelic expression²⁶, including *var* genes in malaria antigenic variation⁸ and developmental activation of highly expressed genes in other differentiated eukaryotic cells. □

Methods

Trypanosomes and immunofluorescence

Trypanosoma brucei bloodstream-form (antigenic type 1.2, MITat 1.2, clone 221a) and procyclic-form strain 427 was used in this work¹. The bloodstream trypanosome transgenic cell line expressing the T7 RNA polymerase and the tetracycline repressor (single marker cell line)²⁰ was used as an initial cell line for the GFP tagging of the ES. We carried out immunofluorescence experiments on cells fixed in methanol at -20°C as described²⁷. Anti-RNA pol I monoclonal (16B1a) and rabbit polyclonal (affinity-purified) antibodies were developed against a His-fusion polypeptide from the large subunit of RNA pol I (residues 1,365–1,521). Polyclonal antisera were further purified over a CNBr–Sephacryl (Pierce) fusion protein column according to the manufacturer's instructions. Immunofluorescence using these antibodies was carried out in 300 mM NaCl and 10% blocking reagent (Roche) in PBS/0.3% Tween 20. As secondary antibodies, we used anti-mouse and anti-rabbit conjugated with Alexa 488 or 596 (Molecular Probes). Images were captured by a Leica DMBRE microscope ($\times 100$ lens) with narrow band filters (Leica) and a CCD camera (Princeton) using IPLab software, and pseudo-coloured and merged in Adobe Photoshop.

We performed DNA FISH using a bacterial artificial clone (BAC) clone (40D16) of ~ 65 kb as a probe, which extended from the 50-base-pair (bp) repeats to ESAG3 of 221ES, and was labelled with digoxigenin (Roche), under described conditions²⁸. RNA FISH was performed using a probe containing the region between 18S and 28S of the rDNA (2.1-kb *HindIII*–*EcoRI*), as described²⁸, except that before fixation the cells were permeabilized in 0.5% Triton X-100, 300 mM sucrose, 100 mM NaCl, 5 mM MgCl_2 , 10 mM PIPES, 2 mM EGTA (pH 6.8) for 5 min at 4°C .

GFP–LacI repressor tagging of the ES

We generated two tagging constructs to mark the active 221ES and the inactive 121ES. For targeting sequences, we used the upstream regions of both promoters: for the 121 ES, a 578-bp *SpeI*–*HindIII* fragment²⁹; and for the 221ES, a 594-bp *SwaI*–*HindIII* fragment⁷. These targeting sequences were cloned upstream of the 256 *lac* operator repeats (10.1-kb *HindIII*–*SmaI* fragment from pAFS59; ref. 3), which were cloned upstream of a 379-bp *HpaI*–*PstI* restriction fragment containing the ES promoter from the DES²⁹ driving a bleomycin selectable marker, with an unregulated untranslated region³⁰. These two constructs were inserted by a single crossover upstream of the ES promoter, resulting in tandem repeat ES promoters that mimic the naturally occurring ES promoter duplication described in 30% of ESs²⁹.

In vitro culture of trypanosomes and DNA transfection procedures have been described^{29,30}. Clonal cell lines were selected and subsequently grown in the presence of $40\text{ }\mu\text{g ml}^{-1}$ phleomycin (Sigma) for the active 221ES targeting experiment, and $1\text{ }\mu\text{g ml}^{-1}$ for the inactive 121ES targeting, as described^{29,30}. Southern analysis confirmed the proper insertion of the cassette (data not shown). To express the GFP–LacI fusion in a tetracycline-dependent manner in the Lac-operator-tagged cell line, we used pLew100 (ref. 20) containing the streptomycin acetyl transferase (SAT) gene instead of bleomycin as a selectable marker, and the EGFP (Clontech) fusion with the LacI (containing SV40nls) from pAFS144 (ref. 3) instead of luciferase. GFP–LacI was detected after induction using an anti-GFP antibody. The levels of GFP–LacI expression varied between cells even in the absence of chromosomal *lac* operator repeats, those cells with lower levels ($\sim 10\%$) were most useful for colocalization studies.

Transcription in permeabilized cells and DNase I treatment

Nascent RNA Br-UTP labelling was essentially done as described¹⁷. Trypanosomes were washed in medium without fetal calf serum and then in transcriptional buffer (50 mM HEPES (pH 7.9), 100 mM KCl, 5 mM MgCl_2 , 5 mM dithiothreitol (DTT), $10\text{ }\mu\text{g ml}^{-1}$ Leupeptin, $80\text{ }\mu\text{g ml}^{-1}$ RNA inhibitor (Roche), 0.5 mM EGTA). Afterwards, Saponin (Calbiotech) was added to a final concentration of $50\text{ }\mu\text{g ml}^{-1}$ (10^7 cells ml^{-1}), and samples were mixed by inversion and incubated for 3 min at room temperature. Cells were washed and resuspended at the same density in transcription buffer, and α -amanitin (Sigma) was added to $100\text{ }\mu\text{g ml}^{-1}$. We next added a mixture of ribonucleoside triphosphates (Roche; final concentration 2 mM ATP, 1 mM CTP, 1 mM GTP) and 0.5 mM Br-UTP (Sigma) to the cell suspension in transcription buffer. Samples were incubated at 33°C in a water bath for 15 min, and the reaction was stopped by fixing the cells with methanol at -20°C . Br-UTP-labelled RNA was detected using a monoclonal anti-BrdUTP (Biocell). Nineteen per cent of the nuclei showed an additional extranucleolar BrUTP focus. The double-labelling experiments showed either closely associated (partially overlapping) or co-localized, but not separate individual signals, except when the inactive 121ES was tagged (Fig. 4e). The DNase I experiment was carried out using the transgenic cell line with the *lac* operator in the 221ES (same as in Fig. 4d), and expression of the GFP–Lac repressor was achieved by inducing cells for 16 h with $1\text{ }\mu\text{g ml}^{-1}$ tetracycline. The DNase I (GibcoBRL) treatment ($1,270\text{ U ml}^{-1}$ for 10 min at room temperature in 10 mM Tris–HCl (pH 7), 10 mM MgCl_2 , 1 mM DTT) was performed on cells fixed with methanol as above. After the cells were washed three times for 15 min in PBS/0.5% Tween 20, the simultaneous immunofluorescence detection of pol I and GFP was carried out as above. No DNA was detected in the nuclei after this treatment, as revealed by the lack of staining with 4',6-diamidino-2-phenylindole dihydrochloride (DAPI). In the control sample (induced cells without DNase I treatment), about 20% of nuclei showed the GFP dot.

Received 19 September; accepted 16 October 2001.

1. Cross, G. A. M. Identification, purification and properties of variant-specific glycoprotein antigens constituting the surface coat of *Trypanosoma brucei*. *Parasitology* **71**, 393–417 (1975).
2. Borst, P. & Ulbert, S. Control of VSG gene expression sites. *Mol. Biochem. Parasitol.* **114**, 17–27 (2001).
3. Robinett, C. C. *et al.* In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *J. Cell Biol.* **135**, 1685–700 (1996).
4. Pays, E., Lips, S., Nolan, D., Vanhamme, L. & Perez-Morga, D. The VSG expression sites of *Trypanosoma brucei*: multipurpose tools for the adaptation of the parasite to mammalian hosts. *Mol. Biochem. Parasitol.* **114**, 1–16 (2001).
5. Bitter, W., Gerrits, H., Kieft, R. & Borst, P. The role of transferrin-receptor variation in the host range of *Trypanosoma brucei*. *Nature* **391**, 499–502 (1998).
6. Pays, E. *et al.* The genes and transcripts of an antigen expression site from *T. brucei*. *Cell* **57**, 835–845 (1989).
7. Zomerdijk, J. C. B. M. *et al.* The promoter for a variant surface glycoprotein gene expression site in *Trypanosoma brucei*. *EMBO J.* **9**, 2791–2801 (1990).
8. Borst, P. & Chaves, I. Mono-allelic expression of genes in simple eukaryotes. *Trends Genet.* **15**, 95–96 (1999).
9. Robinson, N. P., Burman, N., Melville, S. E. & Barry, J. D. Predominance of duplicative VSG gene conversion in antigenic variation in African trypanosomes. *Mol. Cell. Biol.* **19**, 5839–5846 (1999).
10. Roditi, I. *et al.* Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of *Trypanosoma brucei*. *J. Cell Biol.* **108**, 737–746 (1989).
11. Kooter, J. M. & Borst, P. Alpha-amanitin-insensitive transcription of variant surface glycoprotein genes provides further evidence for discontinuous transcription in trypanosomes. *Nucleic Acids Res.* **12**, 9457–9472 (1984).
12. Rudenko, G., Bishop, D., Gottesdiener, K. & Van der Ploeg, L. H. T. Alpha-amanitin resistant transcription of protein coding genes in insect and bloodstream form *Trypanosoma brucei*. *EMBO J.* **8**, 4259–4263 (1989).
13. Rudenko, G., Lee, M. G.-S. & Van der Ploeg, L. H. T. The PARP and VSG genes of *Trypanosoma brucei* do not resemble RNA polymerase II transcription units in sensitivity to Sarkosyl in nuclear run-on assays. *Nucleic Acids Res.* **20**, 303–306 (1992).
14. Laufer, G., Schaaf, G., Bollgott, S. & Günl, A. In vitro analysis of alpha-amanitin-resistant transcription from the rRNA, procyclic acidic repetitive protein, and variant surface glycoprotein gene promoters in *Trypanosoma brucei*. *Mol. Cell. Biol.* **19**, 5466–5473 (1999).
15. Scheer, U. & Hock, R. Structure and function of the nucleolus. *Curr. Opin. Cell Biol.* **11**, 385–390 (1999).
16. Hartshorne, T. & Agabian, N. RNA B is the major nucleolar trimethylguanosine-capped small nuclear RNA associated with fibrillarin and pre-rRNAs in *Trypanosoma brucei*. *Mol. Cell. Biol.* **13**, 144–154 (1993).
17. Pombo, A. *et al.* Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *EMBO J.* **18**, 2241–2253 (1999).
18. Zomerdijk, J. C. B. M., Kieft, R. & Borst, P. Efficient production of functional mRNA mediated by RNA polymerase I in *Trypanosoma brucei*. *Nature* **353**, 772–775 (1991).
19. Chaves, I. *et al.* Subnuclear localization of the active variant surface glycoprotein gene expression site in *Trypanosoma brucei*. *Proc. Natl Acad. Sci. USA* **95**, 12328–12333 (1998).
20. Wirtz, E., Leal, S., Ochatt, C. & Cross, G. A. M. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **99**, 89–101 (1999).
21. Chaves, I., Rudenko, G., Dirks-Mulder, A., Cross, M. & Borst, P. Control of variant surface glycoprotein gene-expression sites in *Trypanosoma brucei*. *EMBO J.* **18**, 4846–4855 (1999).
22. Vanhamme, L. *et al.* Differential RNA elongation controls the variant surface glycoprotein gene expression sites of *Trypanosoma brucei*. *Mol. Microbiol.* **36**, 328–340 (2000).
23. Navarro, M., Cross, G. A. & Wirtz, E. *Trypanosoma brucei* variant surface glycoprotein regulation involves coupled activation/inactivation and chromatin remodeling of expression sites. *EMBO J.* **18**, 2265–2272 (1999).
24. Grosveld, F. Activation by locus control regions? *Curr. Opin. Genet. Dev.* **9**, 152–157 (1999).
25. Cook, P. R. The organization of replication and transcription. *Science* **284**, 1790–1795 (1999).
26. Ohlsson, R., Tycko, B. & Sapienza, C. Monoallelic expression: 'there can only be one'. *Trends Genet.* **14**, 435–438 (1998).
27. Matthews, K. R. & Gull, K. Evidence for an interplay between cell cycle progression and the initiation of differentiation between life cycle forms of African trypanosomes. *J. Cell Biol.* **125**, 1147–1156 (1994).
28. Ersfeld, K. & Gull, K. Partitioning of large and minichromosomes in *Trypanosoma brucei*. *Science* **276**, 611–614 (1997).
29. Navarro, M. & Cross, G. A. DNA rearrangements associated with multiple consecutive directed antigenic switches in *Trypanosoma brucei*. *Mol. Cell. Biol.* **16**, 3615–3625 (1996).
30. Navarro, M. & Cross, G. A. In situ analysis of a variant surface glycoprotein expression-site promoter region in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **94**, 53–66 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank A. F. Straight for the GFP–LacI tagging constructs; E. Wirtz and C. Ochatt for pLew100 and the SAT-derived construct; M. Hoek and G. A. M. Cross for the BAC clone; G. Pierron for the fibrillarin (P2G3) monoclonal antibody; and K. E. Sawin for the anti-GFP rabbit polyclonal antibody. We are grateful to K. Ersfeld for FISH protocols, and D. Robinson for technical advice. We thank A. Baines for technical assistance at the initial stage of this work, and all members of the Gull laboratory for discussions. This work was founded by the Wellcome Trust and the BBSRC.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.N. (e-mail: miguel.navarro@ipb.csic.es).

Thymus medulla consisting of epithelial islets each derived from a single progenitor

Hans-Reimer Rodewald*, Sabine Paul*, Corinne Haller†, Horst Bluethmann‡ & Carmen Blum*

* Department for Immunology, University of Ulm, Albert-Einstein-Allee 11, D-89081 Ulm, Germany

† Basel Institute for Immunology, Grenzacherstrasse 487, CH-4005 Basel, Switzerland

‡ Roche Genetics, F. Hoffmann-La Roche, Grenzacherstr. 124, CH-4070 Basel, Switzerland

The thymus is organized into medullary and cortical zones that support distinct stages of T-cell development. The formation of medulla and cortex compartments is thought to occur through invagination of an endodermal epithelial sheet into an ectodermal one at the third pharyngeal pouch and cleft, respectively^{1–5}. Epithelial stem/progenitor cells have been proposed to be involved in thymus development^{6,7}, but evidence for their existence has been elusive. We have constructed chimaeric mice by injecting embryonic stem (ES) cells into blastocysts using ES cells and blastocysts differing in their major histocompatibility complex (MHC) type. Here we show that the MHC class-II-positive medullary epithelium in these chimaeras is composed of cell clusters, most of which derive from either embryonic stem cell or blastocyst, but not mixed, origin. Thus, the medulla comprises individual epithelial 'islets' each arising from a single progenitor. One thymic lobe has about 300 medullary areas that originate from as few as 900 progenitors. Islet formation can be recapitulated after implantation of 'reaggregated fetal thymic organs'⁸ into mice, which shows that medullary 'stem' cells retain their potential until at least day 16.5 in fetal development. Thus, medulla–cortex compartmentalization is established by formation of medullary islets from single progenitors.

To search for signs of thymic epithelial progenitor cell activity in thymus epithelium, we generated chimaeric mice by injecting ES cells (either CBA (MHC class II, I-A^k) or (BALB/c (I-A^d)) into MHC-mismatched blastocysts (C57BL/6 (I-A^b)) (Fig. 1a). Because MHC class II molecules are highly expressed on medullary epithelium⁹, we used the intense expression of class II molecules, together with cytokeratin which is expressed by thymic epithelium¹⁰, to trace the origin of the medullary epithelium to either the ES cell or the blastocyst. A total of 103 CBA–C57BL/6 and 91 BALB/c–C57BL/6 chimaeras were generated. In all chimaeras, the contribution of ES cell versus blastocyst-derived tissue was analysed in skin (an ectodermal tissue), muscle (a mesodermal tissue) and liver (an endodermal tissue) to include tissues from all three germ layers (Fig. 1a). The relative contribution of ES cells and blastocysts to adult tissues was determined for each chimaera by amplifying a microsatellite marker yielding polymerase chain reaction (PCR) fragments specific for ES or blastocyst genomic DNA. Of 103 CBA–C57BL/6 chimaeras, 48 showed ES cell contribution to all three analysed tissues, 12 to two out of three tissues, 6 to only one of three tissues, and 37 showed no ES cell contribution. We used this analysis to select chimaeras with roughly equal contributions of ES cell and blastocyst ('balanced chimaeras') to increase the probability that both ES- and blastocyst-derived tissues would contribute to the thymus epithelium in ontogeny.

Thymi of mice with balanced chimaerism were analysed further by three-colour-histology to compare the expression of I-A^k, I-A^b and cytokeratin as a pan-thymic epithelial marker¹⁰. The location of medullary areas in thymic sections was identified by monoclonal

antibody MTS 10 (ref. 11), which identifies a mixture of cell types in the medulla (Fig. 1b; and see below). With a cytokeratin-specific antibody, the cortex stained as a dense regular network, but the medullary epithelium was more loosely arranged (Fig. 1c). Staining for cytokeratin and I-A^b (Fig. 1d) or I-A^k (Fig. 1e) identified discrete cell clusters co-expressing either I-A^b or I-A^k with cytokeratin.

In the medullary area shown in Fig. 1d–f, two I-A^b-positive islets (Fig. 1d; blue clusters) were separated by one I-A^k-positive epithelial islet (Fig. 1e; yellow cluster). The I-A^b-positive islet shown in the bottom of Fig. 1f was analysed further at higher resolution for cytokeratin, I-A^b and I-A^k expression. Cytokeratin staining identified single medullary epithelial cells (Fig. 1i) expressing high levels of I-A^b (Fig. 1g). The whole cytokeratin-positive medullary epithelium in this islet was I-A^b-positive (Fig. 1h) and I-A^k-negative (Fig. 1i–k). The cytokeratin-negative, I-A^k-positive cells in this

islet were non-epithelial, haematopoietic cells of bone marrow origin, mostly dendritic cells (data not shown). Owing to mixed bone marrow chimaerism, cytokeratin-negative cells expressing I-A^b or I-A^k can be found in both types of islets.

To determine whether epithelial medullary islets are frequent or rare structures, we searched the DNA-balanced chimaeras for mice in which the thymic epithelium showed an overall significant, histological side-by-side contribution of ES- and blastocyst-derived tissues. Segregation of cytokeratin-positive cells in medullary areas into I-A^k-positive (CBA ES-cell-derived) or I-A^d-positive (BALB/c ES cell-derived) versus I-A^b-positive (blastocyst-derived) cells was evaluated by histological analysis (Fig. 1). The data, summarized in Table 1, show that the number of ES cells versus blastocyst-derived medullary epithelial cells is highly imbalanced in individual islets of a given thymus. Of 25 areas examined in three informative CBA–C57BL/6 chimaeras, and in one BALB/c–C57BL/6 chimaera, 17 areas contained (by two-dimensional histological analysis) only a single islet that could be assigned clearly to either ES cell or blastocyst origin. Six areas had more than one medullary epithelia

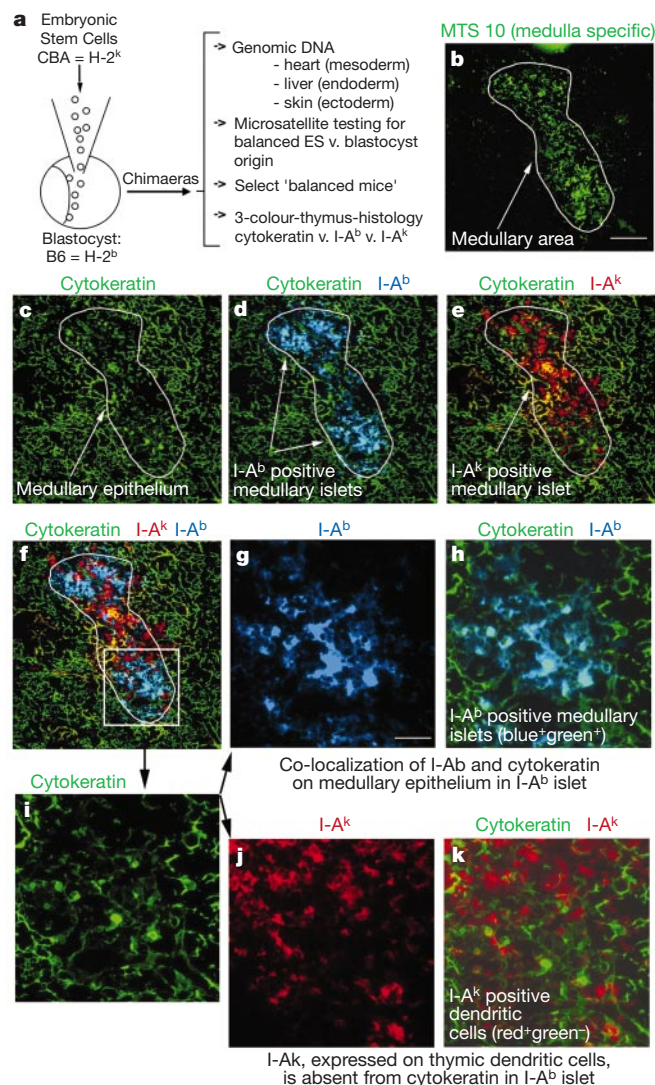


Figure 1 Strategy of the generation and analysis of chimaeric thymic epithelium. **a**, Chimaeric mice were derived by injecting ES cells into MHC-mismatched blastocysts. Chimaerism was determined in heart, liver and skin. **b–k**, Three-colour thymus histology in DNA-balanced chimaeras. Sections were stained with the medulla-specific monoclonal antibody MTS 10 (**b**), cytokeratin (**c–f, h, i, k**; green), I-A^b (**d, f–h**; blue) and I-A^k (**e, f, j, k**; red). One I-A^k-positive epithelial islet (**e**, cells double positive for cytokeratin (green) and I-A^k (red) stain in yellow) is surrounded by two I-A^b-positive islets (**d**, overlay in **f**). Non-epithelial cells of haematopoietic origin are cytokeratin-negative, MHC class-II-positive (**f, k**). Scale bars, 100 μ m (**b**) for **b–f**; 35 μ m (**g**) for **g–k**.

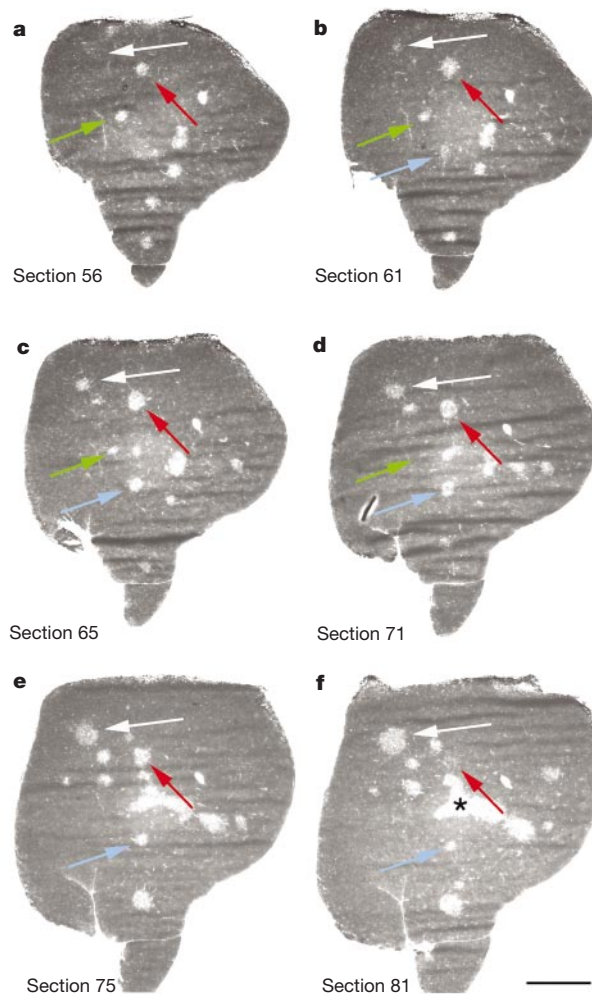


Figure 2 Three-dimensional analysis of medullary islet location in the thymus. One thymus lobe was cut into 300 sections (5 μ m per section), and roughly every fifth section was stained by May–Grünwald–Giemsa to reveal the location of each medullary area. Each smaller area comprises 1–3 islets (Fig. 1, Table 1). **a–f**, Six consecutive slices covering ~125 μ m in depth. Within this depth, some areas appear and expand in size (white arrows), disappear (green arrows), or expand in size and disappear (blue and red arrows). Towards the thymus centre, and in older mice, medullary areas tend to be more confluent and form larger compartments in which the islet character can be lost (asterisk). Scale bar, 1 mm.

origin, and, in these, 10 out of 12 islets could again be assigned to a single origin.

The fact that the origin of MHC class-II-positive, cytokeratin-positive medullary epithelium is highly distinct indicates that the medullary epithelium may comprise individual islets clonally derived from single stem/progenitor cells. The islets vary in diameter from a minimum of $60\text{ }\mu\text{m} \times 40\text{ }\mu\text{m}$ to a maximum of $170\text{ }\mu\text{m} \times 170\text{ }\mu\text{m}$, and have 5–45 epithelial cells in a two-dimensional lattice (Table 1). Longer exposure of the immunofluorescently labelled histological sections revealed MHC class-II-positive, cytokeratin-positive cortical epithelial cells. No correlation was found between the origin of the medullary islet and its surrounding cortex (Table 1).

If medullary epithelium is generated in individual islets rather than as an epithelial sheet, many of these islets may exist 'in isolation' in the cortex rather than as a three-dimensional network. To address this issue, we sectioned a whole thymus lobe into 300 consecutive slices, and stained roughly every fifth section with May–Grünwald–Giemsa to identify the location of each medullary area. One thymus lobe contained about 300 medullary areas, most of which were not connected to one another (Fig. 2). Consecutive sections revealed that smaller areas had a diameter of only about $125\text{ }\mu\text{m}$. It should be noted that the 'islet character' of medullary epithelium was most apparent in the juvenile thymus. Medullary epithelial islets tend to be larger and more confluent in the adult thymus.

These data indicate that medullary epithelium is generated in ontogeny from epithelial stem/progenitor cells rather than from an epithelial cell layer. Such a mechanism of epithelial architecture formation might not only occur during organogenesis, but might also be involved in tissue maintenance or self-reorganization. We

therefore investigated whether medulla–cortex compartmentalization can occur from isolated thymic epithelial cells. Purified epithelium, isolated from day 16.5 of gestation (a stage when the medulla is already well developed), was re-assembled to form epithelial reaggregated fetal thymus organ cultures (RFTOCs)⁸ *in vitro*.

We compared the medulla–cortex architecture, visualized by staining with MTS 10 and the cortex-specific marker MTS 44 (ref. 11), between an endogenous thymus (Fig. 3a, b), RFTOCs maintained *in vitro* in the absence of lymphoid cells (Fig. 3c, d), and RFTOCs grafted under the kidney capsule. In contrast to the normal thymus, RFTOCs maintained *in vitro* lacked the typical medulla–cortex organization (Fig. 3c, d). In RFTOC grafts, medulla–cortex organization was restored (Fig. 3e, f), and indeed resembled the normal thymus (Fig. 3a, b). Stromal cell grafts were colonized by host type, with bone-marrow-derived progenitors developing along normal stages of intrathymic T-cell differentiation, and obeying the rules of positive and negative selection (data not shown).

Because the epithelium had been dissociated completely by enzymatic digestion⁸, 'in vitro reaggregation' occurred from a single-cell suspension yielding a random structure (Fig. 3c, d). Medulla formation in RFTOC grafts might have occurred by the segregation and clustering of pre-existing medullary epithelial cells, or, alternatively, from single progenitors (Fig. 3g). To distinguish between these possibilities, reaggregates were assembled from a 1:1 mixture of C57BL/6 (I-A^b) and BALB/c (I-A^d) fetal thymic epithelium (Fig. 4a). Immunohistological analyses of reaggregate 'tissue' sections to compare the expression of I-A^b and I-A^d showed a random arrangement of the two donor-type epithelia *in vitro* (Fig. 4b).

To avoid any ambiguity caused by colonization of the graft by MHC class-II-positive host (bone-marrow)-derived cells (such as

Table 1 Single epithelial progenitor origin of thymic medullary islets in ES–blastocyst chimaeras

Chimaeras*	Thymus	Medulla			Origin and number of islet cells§		Surrounding cortex‡
		Area	Clonal islet	Islet diameter (μm^2)†	I-A ^b	I-A ^k	
CBA/B6	8B	1	1	170 × 170	45	0	b
		2	1	160 × 90	20	1	b
			2	60 × 40	1	5	b
		3	1	140 × 100	27	0	b
		4	1	150 × 80	28	0	b
		5	?	180 × 150	15	5	Mixed
		6	?	120 × 110	11	4	Mixed
			1	90 × 60	1	7	Mixed
		7	1	110 × 60	0	9	Mixed
	12B	1	1	60 × 40	9	0	b
		2	1	110 × 90	17	0	b
		3	1	90 × 70	14	0	b
		4	1	75 × 80	0	17	Mixed
	44B	1	1	120 × 100	1	26	Mixed
		2	1	115 × 110	28	1	Mixed
			1	90 × 50	2	7	Mixed
			2	80 × 60	13	0	Mixed
		4	1	80 × 80	14	2	Mixed
		5	1	75 × 50	0	6	Mixed
		6	1	80 × 60	1	8	Mixed
			2	90 × 80	12	0	Mixed
		7	1	60 × 40	6	0	Mixed
		8	1	60 × 40	5	0	Mixed
		9	?	80 × 60	7	3	Mixed
BALB/B6	76A	1	1	240 × 50	20	2	Mixed
			2	200 × 70	2	21	Mixed
		2	1	110 × 70	8	3	Mixed
		3	1	130 × 80	19	0	Mixed
		4	1	110 × 60	1	13	Mixed
			?	80 × 80	9	6	Mixed
		5	1	60 × 50	0	5	Mixed

* Data are shown for both CBA→C57BL/6 (CBA/B6, top) and BALB/c→C57BL/6 (BALB/B6, bottom) chimaeras. Chimaeras were analysed at two weeks old for genomic chimaerism. Subsequently, thymy were selected histologically for balanced (CBA/B6, I-A^k versus I-A^b; BALB/B6, I-A^d versus I-A^b) origin of thymic epithelium.

† Islet diameter was measured with Openlab software. For each islet, the origin of thymic epithelium was determined by three-colour histology for I-A^k versus I-A^b versus cytokeratin, or I-A^d versus I-A^b versus cytokeratin, as in Fig. 1. Bold numbers highlight the imbalance of epithelial cells of each origin in individual islets.

‡ The cortex surrounding each islet was examined histologically for matching or non-matching epithelial origin.

§ The origin of medullary epithelial islets was examined by three-colour histology for I-A^k versus I-A^b versus cytokeratin, or I-A^d versus I-A^b versus cytokeratin. Numbers of cells per islet in a two-dimensional view were determined by counting cells co-expressing the indicated MHC class II together with cytokeratin.

dendritic cells), the mixed-MHC reagggregates were implanted into MHC class II^{-/-} C57BL/6 mice, which lack both I-A and I-E molecules¹². The MHC class-II-positive graft was functional *in vivo*, as it promoted the development of mature CD4⁺CD8⁺ thymocytes (Fig. 4c). These thymocytes develop in normal mice, but not in the endogenous class-II^{-/-} thymus¹²⁻¹⁴ (Fig. 4c).

After implantation, the grafts were characterized by well developed medullary areas embedded in the cortex, as revealed by serial tissue sections stained with anti-Thy-1 (Fig. 4d) and MTS 10 (Fig. 4e). The origin of the two medullary areas depicted in Fig. 4e was identified by double staining with anti-I-A^b (Fig. 4f) and anti-I-A^d (Fig. 4g) (Fig. 4h, overlay). Notably, the medullary islet denoted as M₁ was I-A^b-negative but I-A^d-positive (Fig. 4f-h). Conversely,

the medullary islet M₂ was I-A^b-positive but I-A^d-negative (Fig. 4f-h). Sections from reaggregate grafts had about five medullary islets, all of which were either of I-A^b or of I-A^d origin, but not mixed.

The medullary islets in RFTOC grafts (implanted into class II^{-/-} hosts) comprised class-II-positive thymic epithelial cells, as directly shown by double staining for MHC class II and cytokeratin (Fig. 4i-k). Thus, 'clone members' in medullary islets are MHC class-II-positive, cytokeratin-positive cells. However, the medulla contains other cellular phenotypes such as MTS-10-positive, MHC class-II-negative, or cytokeratin-positive, MHC class-II-negative cells (Fig. 4l-o). At present, we lack allelic markers specific for these cells to analyse whether they are also 'clone members'. Collectively, the random epithelial arrangement of the two donor-type epithelia (Fig. 4b) gave rise to medullary islets derived from mutually exclusive epithelia. These data strongly support the findings of medullary islet formation in the normal thymus (Fig. 1).

Normal thymus architecture is dependent on normal development of T cells¹⁵, and medullary epithelium is disorganized completely in the absence of thymocytes¹⁶. We therefore implanted RFTOCs into mice in which T-cell development was blocked. Very early T-cell progenitors present in a T-cell receptor rearrangement-incompetent recipient (RAG-2^{-/-}, MHC class II^{-/-}) were sufficient for the development of medullary islets (data not shown). Collectively, the presence of CD3⁺CD4⁺CD8⁺, rather than mature CD3⁺CD4⁺CD8⁺ and CD3⁺CD4⁺CD8⁺ thymocytes, drives medullary islet formation.

Together, the MHC-mismatched ES-blastocyst chimaeras and the reaggregate grafts suggest a mechanism of thymus organogenesis. The view that thymic structure is formed through invagination of an endodermal into an ectodermal epithelial sheet at the third pharyngeal pouch and cleft¹⁻⁵ is an enduring, well accepted concept. In addition, immunohistological studies suggest the existence of putative 'thymic stem cells' that co-express medullary and cortical markers, and the involvement of such epithelial stem cells in thymus development has been proposed^{6,7}. We now provide evidence for the existence of thymus medulla epithelial stem/progenitor cells. Each progenitor generates one medullary islet, and, on average, 1-3 islets form a medullary area. The medulla-forming cell might possess stem cell characteristics, exemplified by the generation of clonal islets in reagggregates assembled from a single-cell suspension (Fig. 4).

For proof, however, epithelial stem cells need to be purified before self-renewal and precursor-product relationships can be analysed. Our data are also not incompatible with a model in which cell clusters are generated by expansion of a more mature epithelial cell type. Thymus architecture can be strongly affected by various immunosuppressive treatments, for example, irradiation, steroids or cyclosporin A. Moreover, thymic structure deteriorates during certain infections and with age. Whether epithelial progenitors establish the thymic architecture only once during ontogeny, actively renew medullary islets throughout adult life, or are perhaps quiescent until 'reactivated' by thymic injury and subsequent regeneration¹⁷ remains to be determined.

RFTOC transplantations also show that purified thymic epithelium has the remarkable capacity to self-reorganize into a structurally and functionally competent microenvironment promoting T-cell development *in vivo*. Clonality of epithelial islets may result in genetic heterogeneity in the female medulla because the presentation of antigens encoded by the X chromosome might alternate from islet to islet. Whether X-chromosome inactivation in female thymic epithelium, as proposed for female antigen-presenting cells¹⁸, perturbs self-tolerance awaits further experimentation. Our data imply that epithelial stem/progenitor cells are not involved solely in tissue regeneration, but that tissue-specific stem/progenitor cells¹⁹ can generate a specialized epithelial compartment locally. Thymus morphogenesis is apparently dependent on such rare stem/progenitor cells.

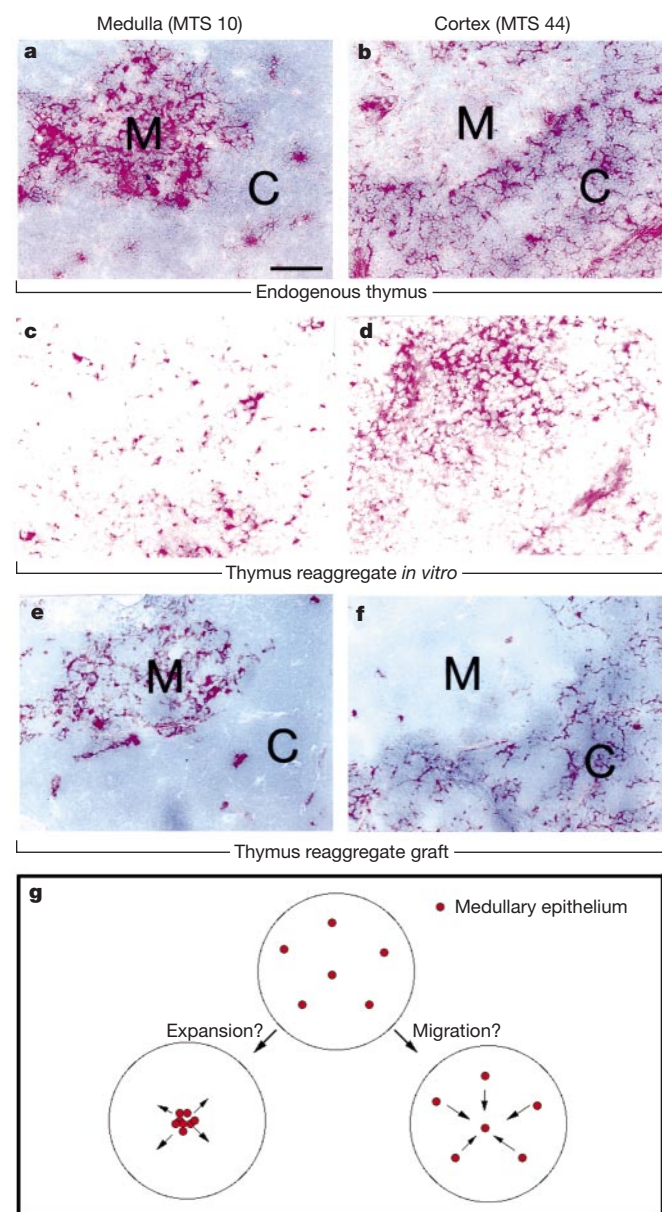


Figure 3 Medulla-cortex re-organization after implantation of thymus epithelial grafts *in vivo*. **a-f**, Tissue sections from normal thymus (**a, b**), RFTOCs maintained *in vitro* (**c, d**) or after transplantation under the kidney capsule (**e, f**) were stained with MTS 10 (**a, c, e**) or with MTS 44 on serial sections (**b, d, f**) to visualize the medulla-cortex organization. Thymocytes, lacking in epithelial RFTOC *in vitro*, were visualized by counterstaining with haematoxylin and eosin (light blue in **a, b, e, f**). Scale bar, 100 μ m. C, cortex; M, medulla. **g**, Medulla-cortex re-organization might occur through the migration ('sorting out') of medullary epithelium, or through the expansion of epithelial progenitors.

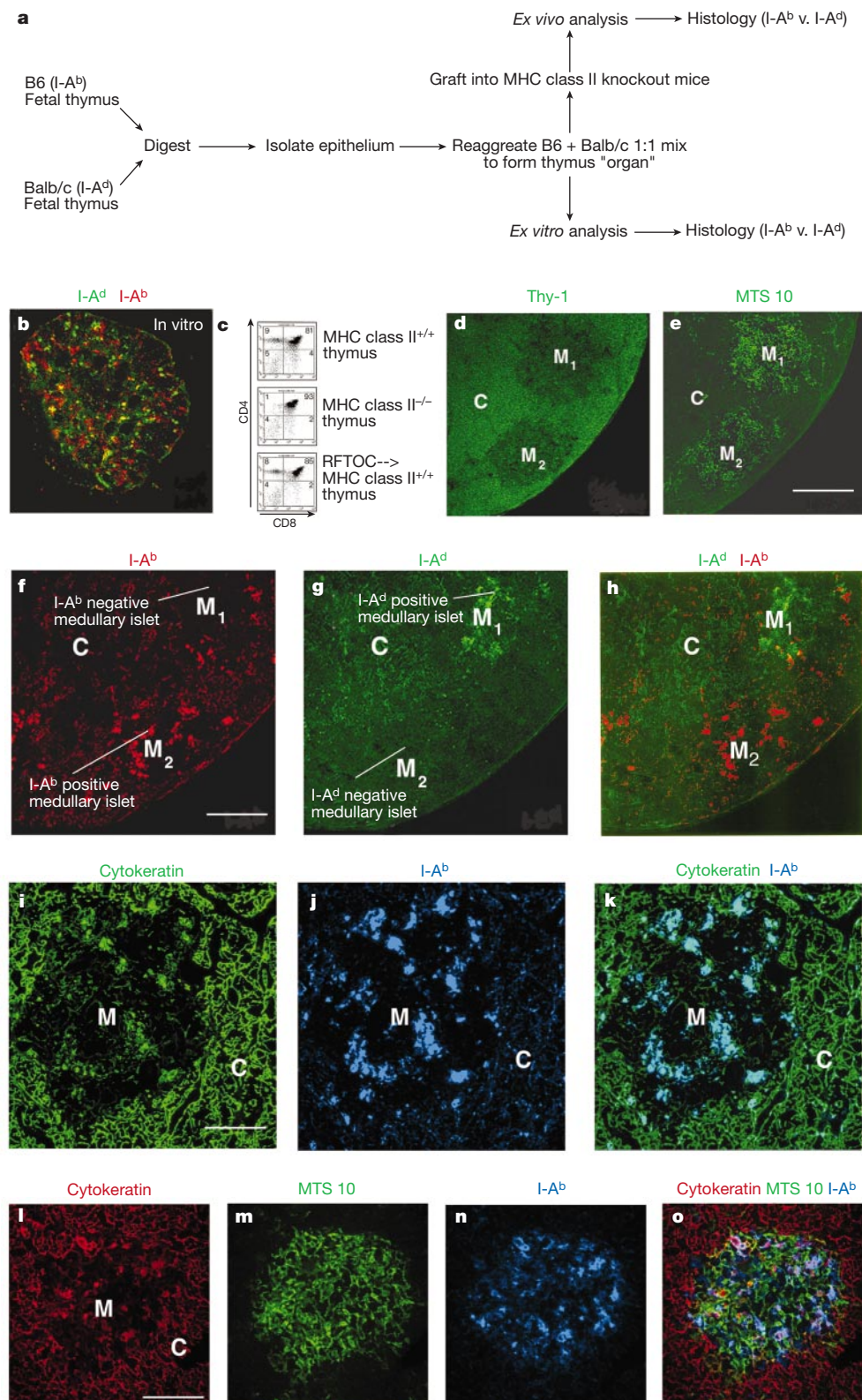


Figure 4 Medullary islets arise in thymus stromal cell grafts generated from two different donor epithelia from either of the two (but not both) donor origins. **a**, Experimental strategy. **b**, Reaggregates maintained *in vitro* for 3 d analysed as tissue sections by immunofluorescence with anti-I-A^b (red) and anti-I-A^d (green). **d–k**, Immunofluorescence histology of reaggregate grafts 4 months after transplantation under the kidney capsule of MHC class-II-deficient hosts. **c**, FACS analysis for expression of CD4 compared with CD8 in wild-type thymus (top), endogenous MHC class II^{-/-} thymus (middle), or MHC class-II⁺ mixed grafts implanted into MHC class II^{-/-} mice (bottom). Grafts were stained with

anti-Thy-1 (**d**, cortex and medulla densely and weakly stained, respectively), with MTS 10 (**e**), or with anti-I-A^b (**f**) or with anti-I-A^d (**g**) (overlay in **h**). The medullary islet M₁ is I-A^b-negative, I-A^d-positive; islet M₂ is I-A^b-positive, I-A^d-negative (**f–h**). **i–k**, Co-localization (**k**) of cytokeratin (**i**) and I-A^b (**j**) in a medullary islet in RFTOC grafts in a MHC class-II^{-/-} hosts. **l–o**, Three-colour histology for cytokeratin (**l**), MTS10 (**m**), I-A^b (**n**), and overlay of all colours (**o**) in C57Bl/6 mice. Scale bars, 500 μ m (**e** for **b**, **d**, **e**; 375 μ m (**f** for **f–h**; 200 μ m (**i** for **i–k**; 100 μ m (**l** for **l–o**.

Methods

Generation of ES–blastocyst chimaeras

To generate MHC-mismatched chimaeric mice, either CBA/CaOlaHsd (H-2^k), or BALB/c (H-2^d) ES cells (both from Eurogentec) were injected into C57BL/6 blastocysts as described¹². All chimaeric mice were analysed at 2 weeks of age. From each chimaera, thymic lobes were frozen for subsequent histological analysis, and genomic DNA was prepared from liver, muscle and skin. Screening several oligonucleotide pairs (MapPairs, Research Genetics) allowed us to identify two pairs (D8Mit294 for CBA versus C57BL/6; D8Mit293 for BALB/c versus C57BL/6) that generated a robust polymorphic size difference when comparing CBA or BALB/c with C57BL/6 genomic DNA. We quantified DNA chimaerism by comparing the two bands diagnostic for each tissue with parallel PCR analysis of control templates with defined ratios of DNA from the two strains. The contribution of ES versus blastocyst origin to the generation of liver, muscle and skin was determined for each chimaeric mouse.

Monoclonal antibodies, histology and flow cytometry

For flow cytometry²⁰, we used phycoerythrin-labelled H129.19 (anti-CD4) and biotin-53-6.7 (anti-CD8) (both from Gibco-BRL). Streptavidin-APC (Molecular Probes) was used as second-step reagent. Thymic tissue (chimaeric lobes, or RFTOC kept *in vitro*, or grafted RFTOC) was embedded in OCT medium (Sakura Finetek), and snap frozen, and 5-µm sections were cut with a cryostat. Sections were air dried for 60 min, acetone fixed for 8 min, and air dried overnight. Sections were then rehydrated in PBS/1% BSA. Monoclonal antibodies diluted in PBS/1% BSA were added directly onto the sections and incubated for 60 min. The following monoclonal antibodies were used: MTS 10 (ref. 11; rat IgM; anti-medullary epithelium); MTS 44 (anti-cortical epithelium; ref. 11), Cy3.5-labelled (using Amersham Life Science kit) or unlabelled anti-multi-cytokeratin (clone C11, mouse IgG1, 1:100 diluted; Neomarkers), biotinylated anti-I-A^b (clone AF6-120.1, 1:50 diluted; used for I-A^b versus I-A^d), biotinylated anti-I-A^b (clone 25-9-17, 1:50 diluted; used for I-A^b versus I-A^k), fluorescein isothiocyanate (FITC)-labelled or unlabelled anti-I-A^d (clone AMS-32.1, mouse IgG2b, 1:10 diluted), unlabelled anti-I-A^k (clone 11-5.2, mouse IgG2b, 1:50 diluted) (all from Pharmingen). Second-step reagents for immunofluorescence were: streptavidin-labelled Texas red (1:100 diluted; Caltag), Cy5-labelled streptavidin (1:100 diluted; Amersham-Pharmacia), FITC-labelled anti-mouse-IgG2b (1:50), FITC-anti-rat-IgM (1:50) (both Pharmingen), FITC-labelled-anti-mouse-IgG1 (1:50), Texas red anti-mouse-IgG2b (1:200) (both Southern Biotechnology). After being washed, the sections were mounted in Fluoromount (Southern Biotechnology). We recorded images on a Zeiss Axioskop with an ORCA ER camera (Hamamatsu), and processed them using Openlab software (Improvision). Immunohistochemistry (Fig. 3) was done as described²⁰.

Reaggregation and tissue transplantation

Thymus stroma reaggregates were prepared from fetal thymus stroma cells as described⁸. Briefly, pooled thymi from day 16.5 fetal mice were digested with Collagenase-Dispase (1 mg ml⁻¹; Boehringer Mannheim) and DNase I (0.5 mg ml⁻¹; Sigma) in PBS/5% fetal calf serum (FCS). The tissue was digested for a total of 60 min with rotation at 37 °C. During this time, the Collagenase-Dispase-DNase I mix and the released cells were removed after 30 min, and fresh Collagenase-Dispase-DNase I mix was added for another 30 min. To remove haematopoietic cells, we incubated this thymus cell suspension with anti-pan-CD45 monoclonal antibody (clone I 3/2; Sigma) at 10 µg ml⁻¹ in PBS/5% FCS for 30 min on ice, washed the cells twice in PBS/5% FCS, and removed the CD45⁺ cells in two rounds by depletion with sheep anti-mouse-Ig magnetic beads (Dynabeads) at a ratio of ten beads to one cell. The resulting cell population was highly enriched for thymic epithelium.

To generate reaggregates, we spun epithelial cells into a firm cell pellet at 1,280g for 5 min, completely removed the supernatant, and resuspended the cells into a slurry by adding 0.5–1 µl of PBS/5% FCS. The slurry was drawn into a capillary, and little droplets of ~200 nl (corresponding to ~5 × 10⁵ stromal cells) were placed as 'standing drops' onto the dry surface of a nucleopore filter floating in a 10-cm petri dish filled with 5 ml of DMEM with 10% FCS. After 3–5 d, the solidified reaggregates were retrieved from the filters, and either snap frozen for histological analysis, or implanted into mice. Thymus grafting was done as described²¹.

Received 23 July; accepted 11 October 2001.

1. Crisan, C. Die Entwicklung des thyreo-parathyreo-thymischen Systems der weissen Maus. *Z. Anat. Entwicklungsgesch.* **104**, 327–358 (1935).
2. Norris, E. H. The morphogenesis and histogenesis of the thymus gland in man: in which the origin of the Hassall's corpuscles of the human thymus is discovered. *Contrib. Embryol.* **166**, 191–221 (1938).
3. Cordier, A. C. & Haumont, S. M. Development of thymus, parathyroids, and ultimo-branchial bodies in NMRI and nude mice. *Am. J. Anat.* **157**, 227–263 (1980).
4. von Gaudecker, B. Functional histology of the human thymus. *Anat. Embryol.* **183**, 1–15 (1991).
5. Manley, N. R. Thymus organogenesis and molecular mechanisms of thymic epithelial cell differentiation. *Semin. Immunol.* **12**, 421–428 (2000).
6. Lampert, I. A. & Ritter, M. A. in *Thymus Update* (eds Kendall, M. D. & Ritter, R. M.) 5–25 (Harwood Academic, London, 1988).
7. Ritter, M. A. & Boyd, R. L. Development in the thymus: it takes two to tango. *Immunol. Today* **14**, 462–469 (1993).
8. Anderson, G., Jenkinson, E. J., Moore, N. C. & Owen, J. J. MHC class II-positive epithelium and mesenchyme cells are both required for T-cell development in the thymus. *Nature* **362**, 70–73 (1993).
9. Van Ewijk, W., Rouse, R. V. & Weissman, I. L. Distribution of H-2 microenvironments in the mouse thymus. Immunoelectron microscopic identification of I-A and H-2K bearing cells. *J. Histochem. Cytochem.* **28**, 1089–1099 (1980).

10. Suniara, R. K., Jenkinson, E. J. & Owen, J. J. Studies on the phenotype of migrant thymic stem cells. *Eur. J. Immunol.* **29**, 75–80 (1999).
11. Boyd, R. L. *et al.* The thymic microenvironment. *Immunol. Today* **14**, 445–459 (1993).
12. Kontgen, F., Suss, G., Stewart, C., Steinmetz, M. & Bluethmann, H. Targeted disruption of the MHC class II Aa gene in C57BL/6 mice. *Int. Immunol.* **5**, 957–964 (1993).
13. Grusby, M. J., Johnson, R. S., Papaioannou, V. E. & Glimcher, L. H. Depletion of CD4⁺ T cells in major histocompatibility complex class II-deficient mice. *Science* **253**, 1417–1420 (1991).
14. Gogrove, D. *et al.* Mice lacking MHC class II molecules. *Cell* **66**, 1051–1066 (1991).
15. van Ewijk, W. T-cell differentiation is influenced by thymic microenvironments. *Annu. Rev. Immunol.* **9**, 591–615 (1991).
16. Rodewald, H. R., Brocker, T. & Haller, C. Developmental dissociation of thymic dendritic cell and thymocyte lineages revealed in growth factor receptor mutant mice. *Proc. Natl Acad. Sci. USA* **96**, 15068–15073 (1999).
17. Beschorn, W. E., Di Gennaro, K. A., Hess, A. D. & Santos, G. W. Cyclosporine and the thymus: influence of irradiation and age on thymic immunopathology and recovery. *Cell. Immunol.* **110**, 350–364 (1987).
18. Stewart, J. J. The female X-inactivation mosaic in systemic lupus erythematosus. *Immunol. Today* **19**, 352–357 (1998).
19. Weissman, I. L. Stem cells: Units of development, units of regeneration, units in evolution. *Cell* **100**, 157–168 (2000).
20. Rodewald, H.-R., Ogawa, M., Haller, C., Waskow, W. & DiSanto, J. P. Pro-thymocyte expansion by c-kit and the common cytokine receptor γ chain is essential for repertoire formation. *Immunity* **6**, 265–272 (1997).
21. Rodewald, H.-R., Kretschmar, K., Swat, W. & Takeda, S. Intrathymically expressed c-kit ligand (stem cell factor) is a major factor driving expansion of very immature thymocytes *in vivo*. *Immunity* **3**, 313–319 (1995).

Acknowledgements

We thank V. Bührmann, B. Kugelberg and Y. Lang for technical assistance; F. Melchers, S. Gilfillan, J. Fehling, U. Grawunder and W. Knöchel for discussions; S.-S. Tan for collaborating on early stages related to this work; R. Boyd for antibodies and discussion; and G. Anderson, E. Jenkinson and J. J. T. Owen for an introduction to epithelial reaggregation. H.R.R. was supported by the Deutsche Forschungsgemeinschaft and Landesforschungsschwerpunkt Baden-Württemberg. The Basel Institute for Immunology was founded and supported by F. Hoffmann La Roche, Basel.

Correspondence and requests for material should be addressed to H.R.R. (e-mail: hans-reimer.rodewald@medizin.uni-ulm.de).

c-Myc regulates mammalian body size by controlling cell number but not cell size

Andreas Trumpp^{*,†}, Yosef Refaelli[†], Thordur Oskarsson^{*}, Stephan Gasser^{*}, Mark Murphy^{*}, Gail R. Martin[‡] & J. Michael Bishop[†]

^{*} Swiss Institute for Experimental Cancer Research (ISREC), Ch. des Boveresses 155, CH-1066 Epalinges, Switzerland

[†] The G. W. Hooper Foundation and Department of Microbiology and Immunology, University of California at San Francisco (UCSF), San Francisco, California 94143-0552, USA

[‡] Department of Anatomy and Program in Developmental Biology UCSF, San Francisco, California 94143-0452, USA

Overexpression of the proto-oncogene *c-myc* has been implicated in the genesis of diverse human tumours. *c-Myc* seems to regulate diverse biological processes, but its role in tumorigenesis and normal physiology remains enigmatic¹. Here we report the generation of an allelic series of mice in which *c-myc* expression is incrementally reduced to zero. Fibroblasts from these mice show reduced proliferation and after complete loss of *c-Myc* function they exit the cell cycle. We show that *Myc* activity is not needed for cellular growth but does determine the percentage of activated T cells that re-enter the cell cycle. *In vivo*, reduction of *c-Myc* levels results in reduced body mass owing to multiorgan hypoplasia, in contrast to *Drosophila dmec* mutants, which are smaller as a result of hypotrophy². We find that *dmec* substitutes for *c-myc* in fibroblasts, indicating they have similar biological activities. This suggests there may be fundamental differences in the

mechanisms by which mammals and insects control body size. We propose that in mammals *c-Myc* controls the decision to divide or not to divide and thereby functions as a crucial mediator of signals that determine organ and body size.

To study the normal function of *c-myc* in mice, we used gene targeting to generate a series of new *c-myc* alleles, including two null alleles (*c-myc*^{ΔORF} and *c-myc*^{ΔORFrec}), a hypomorphic conditional allele (*c-myc*^{loxN}) and a conditional allele with apparent wild-type activity (*c-myc*^{loxN}) (Fig. 1a and Supplementary Information Fig. 1). As expected, embryonic day (E) 9.5 embryos homozygous for *c-myc*^{ΔORF} produced no detectable *c-myc* transcript. However, *N-myc* RNA levels were twice as high as normal, whereas messenger RNA levels of other genes in the *myc/max* gene network were normal (Fig. 1b). As previously described³, complete loss of *c-myc* function causes embryonic lethality at ~E10.5. We found that E9.5 *c-myc*^{ΔORF/ΔORF} embryos were severely reduced in size and had pale yolk sacs that apparently lacked primitive erythrocytes (Fig. 1c, and data not shown). Furthermore, whereas yolk sac cells from wild-type embryos produced erythroid- and myeloid-like colonies in methylcellulose culture, null mutant yolk sac cells did not (Fig. 1d). This suggests that *c-myc* is required for primitive haematopoiesis and provides a credible explanation for the observed embryonic lethality⁴.

Mice heterozygous for the null allele (*c-myc*^{ΔORF/+}) had a smaller body mass than their wild-type littermates. This was apparent by E16.5 (data not shown), and newborn *c-myc*^{ΔORF/+} pups could readily be distinguished from wild-type littermates by their small body size (Fig. 2a). Moreover, *c-myc*^{ΔORF/+} pups grew more slowly and never reached the size of their normal littermates (Fig. 2b and Supplementary Information Fig. 2). On average, adult heterozygotes (1 to 8 months) were 18.3% ± 5.4 (mean ± standard error), *n* = 42 (females) to 22.4% ± 5.4, *n* = 104 (males) smaller than wild-type littermate controls.

These observations prompted us to examine body size in mice that express more or less *c-myc* than null heterozygotes. The *c-myc*^{loxN} allele we generated is a hypomorphic allele, owing to the presence of a *Pgk-neo* cassette inserted in its 3' untranslated region (Fig. 1a, e). Using combinations of *c-myc*⁺, *c-myc*^{loxN} and *c-myc*^{ΔORF}, we generated an allelic series of mice (Supplementary Information Fig. 1b) in which *c-Myc* levels were incrementally reduced from normal to zero (Fig. 1b, e): *c-myc*^{+/+} > *c-myc*^{loxN/+} > *c-myc*^{ΔORF/+} ≅ *c-myc*^{loxN/loxN} > *c-myc*^{ΔORF/loxN} > *c-myc*^{ΔORF/ΔORF}. This resulted in an incremental decrease in body mass (Fig. 2b); complete loss of function caused embryonic lethality (Fig. 1c).

All organs from *c-myc*^{ΔORF/+} mice weighed less than normal. However, for some organs (for example, lung, brain, liver) the ratio of their weight to whole body weight was nearly normal. For others, including carcass (muscle, connective tissue, skin, bones), white fat, and all lymphoid organs (thymus, spleen, lymph nodes and bone marrow), the organ/body weight ratio was reduced by 24% to 49% (data not shown). Reduced organ size can be caused by a decrease in cell number, cell size, or both. We found that cell number was lower in the spleen, lymph nodes and bone marrow of *c-myc* mutant mice, and that the diminution correlated with their allelic constitution (Fig. 2c). Forward scatter analysis (FSC) using flow cytometry revealed no significant effect on the size of mutant primary embryo fibroblasts, and cells isolated from mutant lung, gut, heart, kidney and brain (Fig. 3a, and data not shown). In addition, FSC values were similar for naive and activated CD4⁺ and CD8⁺ T cells or various other haematopoietic cell types present in bone marrow and spleen such as megakaryocytes (CD41⁺TER-119⁺), erythroblasts (TER119⁺), granulocytes/macrophages (Gr-1⁺Mac-1⁺) and haematopoietic stem cells (c-kit⁺Thy1.1⁺Lin⁻/Sca-1⁺) (Fig. 4c, e, and data not shown). Furthermore, the number of hepatocytes per microscopic field in sections of liver from mice carrying different allelic combinations was very similar (wild type, 456 ± 23; *c-myc*^{ΔORF/loxN}, 439 ± 16; and MxCre, *c-myc*^{loxN/loxN} in which over 90%

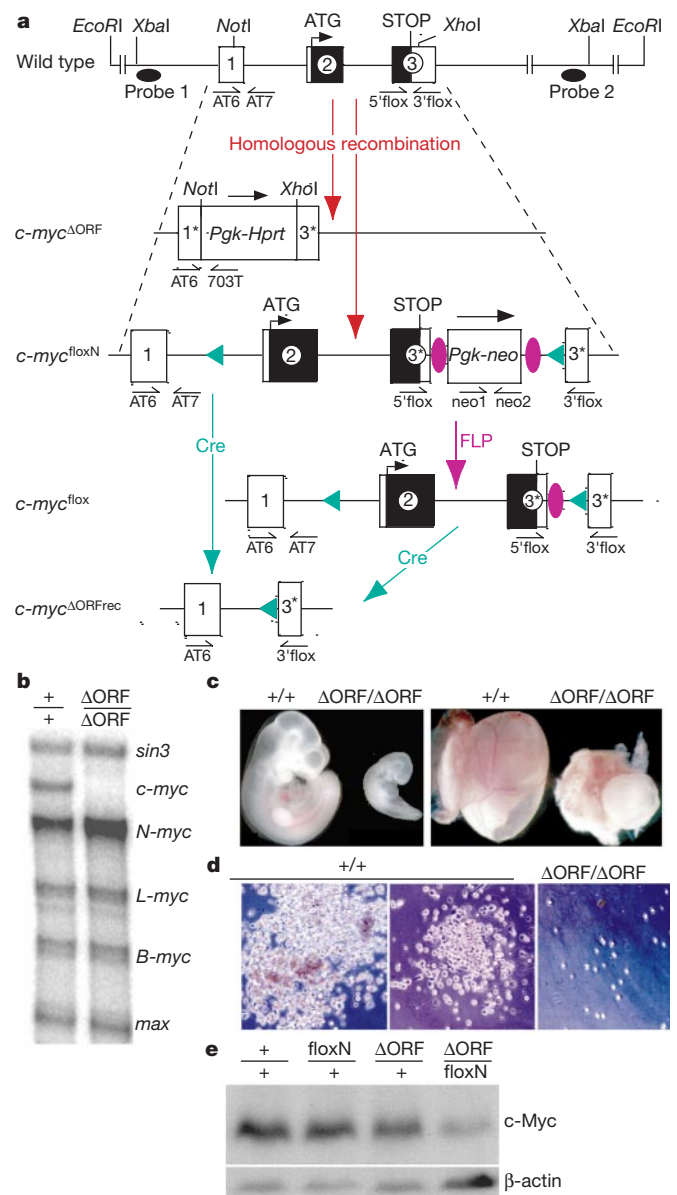


Figure 1 Generation of mutant *c-myc* alleles and analysis of mutant embryos.

a, Schematic representation of the wild-type mouse *c-myc* locus (top). The three *c-myc* exons are illustrated as boxes; regions in black indicate the open reading frame (ORF). Restriction sites are shown on top. Black ovals indicate probes used for Southern blot analysis. Two alleles, *c-myc*^{ΔORF} and *c-myc*^{loxN}, were obtained by homologous recombination. In *c-myc*^{ΔORF} the entire ORF was replaced with a *Pgk-Hprt* minigene²³. The remaining exon sequences are marked 1* and 3*. In *c-myc*^{loxN}, a *Pgk-neo* cassette flanked by *frt* sites, the recombination target sequences for the FLP recombinase (pink ovals), and followed by a *loxP* site (green triangle), was inserted into the 3' untranslated region of the *c-myc* gene. A second *loxP* site was inserted into intron 1. The *c-myc*^{loxN} allele is converted to *c-myc*^{lox} or *c-myc*^{ΔORFrec} by Flp- or Cre-mediated recombination, respectively. Small arrows indicate the direction of transcription of the *c-myc*, *Pgk-Hprt* and *Pgk-neo* genes. Locations of the primers used for genotyping (see Supplementary Information) are indicated underneath each scheme. **b**, Expression analysis of *Myc* family members by RNase protection assays using E9.5 embryo RNA. **c**, Wild-type and mutant E10.25 embryos (left) and yolk sacs (right). **d**, Haematopoietic progenitor colony assays of cells from individual E9.5 yolk sacs cultured in methylcellulose. Examples of erythroid-like (left) and myeloid-like (middle) colonies formed by wild-type yolk sac cells (147 ± 10 and 49 ± 10 per yolk sac, respectively; *n* = 4). No colonies developed from *c-myc*^{ΔORF/ΔORF} yolk sac cells (*n* = 3). **e**, Western blot showing the level of c-Myc protein in activated CD4⁺ thymocytes isolated from mice carrying the *c-myc* alleles indicated.

of liver cells lack *c-myc* function (see Methods), 447 ± 28 ; $n > 3$ for each genotype). Together, these data suggest that *c-myc* mutant mice suffer from multiorgan hypoplasia without any apparent alteration in cell size.

To test whether *c-myc* is required for cell proliferation, we isolated primary mouse embryonic fibroblasts (MEFs) from E13.5 embryos and compared their population growth under identical conditions in culture. MEF population doubling time increased with decreasing levels of *c-myc* expression from 31 h in the wild-type to 55 h in *c-myc*^{ΔORF/floxN} cultures, whereas individual cell size and cell cycle distribution was found to be similar in all cell populations (Fig. 3a and data not shown). Primary MEFs, isolated from E9.5 *c-myc*^{ΔORF/ΔORF} embryos about 24 h before embryos died, flattened and did not proliferate, whereas wild-type cells did (data not shown), indicating that complete loss of *c-myc* function blocks proliferation of E9.5 MEFs.

To study the requirement for c-Myc function in cells that proliferate in culture, we isolated E13.5 *c-myc*^{floxN/floxN} MEFs and infected them with a retrovirus (pMIG-Cre) that co-expresses Cre and GFP from a bicistronic transcript⁵. High-speed cell sorting yielded a green fluorescent protein (GFP)-positive *c-myc*^{floxN/floxN} cell population in which 97.5% ($\pm 1.2\%$) and 99.3% ($\pm 0.4\%$) alleles had been converted into null alleles after 2 and 5 days post sorting, respectively. *c-myc*^{ΔORF/ΔORF} cultures stopped expanding and within 24–48 h after cell sorting, the mutant cells became abnormally flat, resembling normal cells deprived of serum (Fig. 3b, c and data not shown). Most of the null mutant MEFs had a diploid DNA content (data not shown), indicating that they had exited the cell cycle and were arrested in G₀, as in ref. 6.

To determine whether cell proliferation is still *c-myc* dependent in immortalized fibroblasts, we generated 3T9 cells⁷ from wild-type and *c-myc*^{floxN/floxN} embryos. After infection with pMIG-Cre, two cell types appeared, those that expressed GFP and Cre but not *c-myc*,

and those that did not express GFP or Cre and therefore expressed normal levels of *c-myc* (see Methods). Both cell populations were sorted and mixed in approximately equal numbers, and the proportion of each cell type in the cell population was assayed by flow cytometry at two-day intervals. *c-myc*^{ΔORF/ΔORF} cells were at a strong competitive disadvantage to cells that contained a functional *c-myc* gene (Fig. 3d). Ectopic expression of the *Drosophila myc* gene, *dmyc*⁸, before infection with pMIG-Cre enabled *c-myc* null cells to proliferate and compete with cells that expressed both dMyc and c-Myc (Fig. 3d). In none of these growth competition experiments did we detect any significant difference in cell size between GFP⁺ and GFP[−] cells, and neither N-myc nor L-myc expression was detected (see Methods). These data suggest that *c-myc* is crucial for proliferation of immortalized fibroblasts and that *dmyc* can at least partially replace the function of mouse *c-myc*.

To explore the relationship between cell growth (increase in cell mass) and division (increase in cell number) in *c-myc* mutants, we examined primary lymphocytes, in which the two processes can be studied separately. T-cell receptor (TCR) activation in naive peripheral T cells induces *c-myc* and causes a transition from small quiescent T cells to large T-cell blasts (Fig. 4a, our unpublished observation and ref. 9). In addition to the increase in cell size that is caused by an increased rate of biosynthesis and metabolism, antigenic stimulation also up-regulates surface marker expression, including CD25 and CD44. These signs of TCR activation occur within the first 12–24 h of stimulation, before the primed cells leave quiescence and re-enter the G₁ phase of the cell cycle (Fig. 4a). Naive

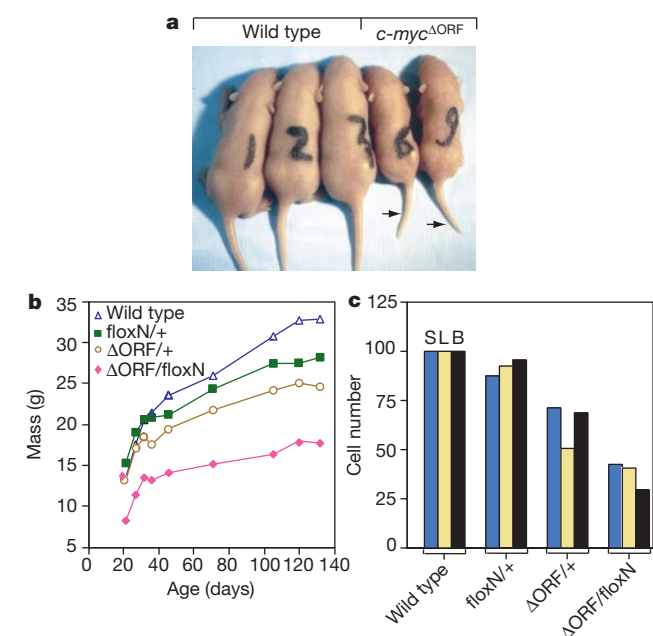


Figure 2 Effect of altering *c-myc* expression on body and organ size. **a**, Wild-type (numbered 1–3) and *c-myc*^{ΔORF/+} (numbered 6 and 9) littermates on postnatal day 5. We note the smaller body size and shorter tail (arrows) of the *c-myc*^{ΔORF/+} pups. **b**, Growth curves of individual female littermates carrying different combinations of the *c-myc*, *c-myc*^{ΔORF} and *c-myc*^{floxN} alleles (*c-myc* allelic series). **c**, Cell number in organs isolated from mice carrying different *c-myc* allelic combinations. For each genotype, the cell number in the spleen (S), lymph nodes (L), and bone marrow (B) was divided by the cell number in that organ from a wild-type littermate, and the ratio is shown. Similar results were obtained for three litters and one typical example is shown.

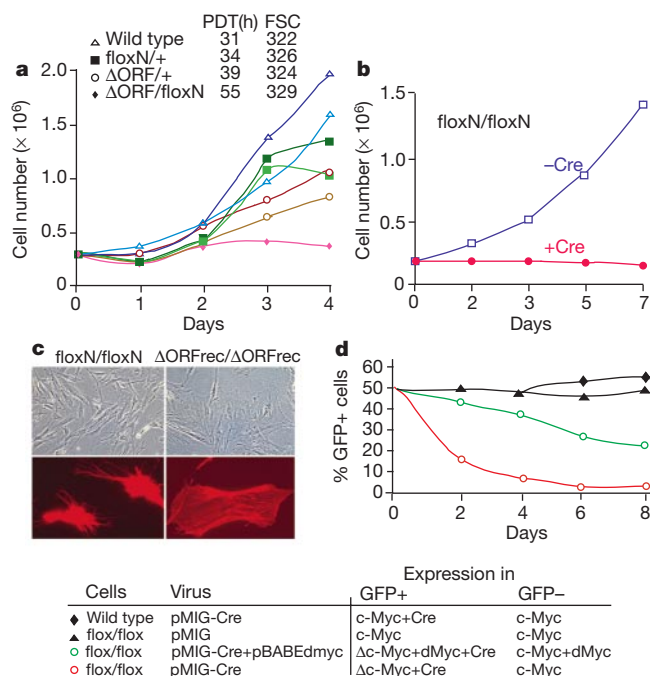


Figure 3 Effect of altering *c-myc* expression on cell proliferation and morphology in primary and immortalized fibroblasts. **a**, Growth curves of primary MEFs (passaged once after isolation, P1) isolated from individual littermates and grown for four days. Population doubling time (PDT) and cell size as estimated by forward scatter analysis (FSC) are shown in the insert. **b**, Growth curve of primary *c-myc*^{floxN/floxN} mouse embryonic fibroblasts (MEFs) after infection with pMIG-Cre and subsequent cell sorting (day 0). **c**, Morphology (top shows the phase contrast and bottom shows the phalloidin stain) of *c-myc*^{floxN/floxN} and *c-myc*^{ΔORF/ΔORF} MEFs. **d**, Competition expansion assay between 3T9 fibroblasts that express c-Myc, Δc-Myc and/or dMyc. Cells and viruses used as well as gene expression in the GFP⁺ and GFP[−] populations are indicated in the table below. The *c-myc* allele that is being expressed is indicated as follows: c-Myc, wild-type or *c-myc*^{floxN} allele; Δc-Myc, *c-myc*^{ΔORF} null allele. The percentage of GFP⁺ cells was determined by fluorescence-activated cell sorting (FACS) analysis at the times indicated. As GFP⁺ cells were initially about 50% of the population (see Methods), all percentages were normalized to 50%.

T-cells from *c-myc*^{ΔORF/floxN} mice increased in size and expressed normal levels of CD25 and CD44, indicating that they had blasted like wild-type cells (Fig. 4b). To determine whether *c-myc* null cells behave similarly, we isolated primary CD4⁺ and CD8⁺ T cells from polyI-polyC treated MxCre; *c-myc*^{ΔORF/flox} mice^{10,11} (see Methods) in which 62% ± 10% of the cells were converted to *c-myc*^{ΔORF/ΔORF}. In response to antigenic stimulation, the resulting *c-myc* null T cells blasted normally (Fig. 4c and data not shown). Neither N-myc nor L-myc expression was detected in normal or *c-myc* mutant T cells (see Methods) so these data indicate that Myc function is not required for cell growth, an essential feature of T-cell activation in response to antigenic stimulation.

To determine whether *c-myc* mutant T-cell blasts re-enter the cell cycle in a normal fashion, we labelled naive T cells with the fluorescent dye carboxyfluorescein-succinimidyl-ester (CFSE) and analysed their proliferative potential following TCR activation. Decreasing *c-myc* levels caused a diminution in proliferative response to primary activation (Fig. 4d). The percentage of T cells that proliferated in response to TCR stimulation decreased from 68.5% in wild-type to 14.9% in *c-myc*^{ΔORF/floxN} cells, with a concomitant increase in the percentage of cells that failed to re-enter the cell cycle (Fig. 4e). The percentage of blasted CD25^{hi} wild-type and *c-myc*^{ΔORF/floxN} cells was similar in these populations (87% versus 86% respectively), further supporting our finding that mutant T cells show a normal activation response. The strong negative correlation between c-Myc levels (Fig. 1e) and the proportion of blasted cells that remain in a resting state suggests an essential and rate-limiting function for c-Myc in promoting the G0 to G1 transition. In addition, we found that low levels of c-Myc through-

out the four-day CFSE assay also resulted in decreased T-cell survival. This raises the possibility that c-Myc may have an additional role in promoting survival of primary T cells (Y.R., L. VanParijs, A.T. and J.H.B., submitted).

It has been suggested that the CDK inhibitor *p27*^{KIP1} (here referred to as *p27*) is a critical cell-cycle target of c-Myc^{12,13}. To study whether the two genes interact, we generated mice carrying various combinations of *p27* (ref. 14) and *c-myc*^{ΔORF} alleles. First, we found that body size was equally reduced when one copy of *c-myc*^{ΔORF} was present in *p27*^{+/-} or *p27*^{-/-} animals (17.8% ± 5.3% *n* = 38), similar to what was observed in comparisons between wild-type (normal *p27*) and *c-myc*^{ΔORF} mice (18.3% ± 5.4%, *n* = 42, see above). Second, *c-myc*^{ΔORF/ΔORF}; *p27*^{-/-} embryos died at mid-gestation and appeared similar to *c-myc* null mutant embryos that were wild type at the *p27* locus (Fig. 1c and data not shown). Third, loss of *p27* function did not alter the proliferative capacity of activated *c-myc*^{ΔORF/+} T cells (Fig. 5). Thus, we were unable to detect a genetic interaction between *c-myc* and *p27*.

Perhaps our most interesting conclusion is that reduction of *c-myc* expression *in vivo* results in an overall decrease in body size. A similar result has been obtained in *Drosophila* by reducing expression of the *c-myc* homologue, *dmyc*. However, when the level of Myc is reduced in *Drosophila*, cell size is reduced (organ hypotrophy), but there appears to be no effect on cell number, indicating that dMyc is essential for cell growth². In contrast, when c-Myc levels are reduced in mice, cell size appears to be normal but cell number is reduced (organ hypoplasia). By studying primary T-cell activation, a process in which it is possible to distinguish between effects on cell growth and cell division, we showed that Myc is not required for the

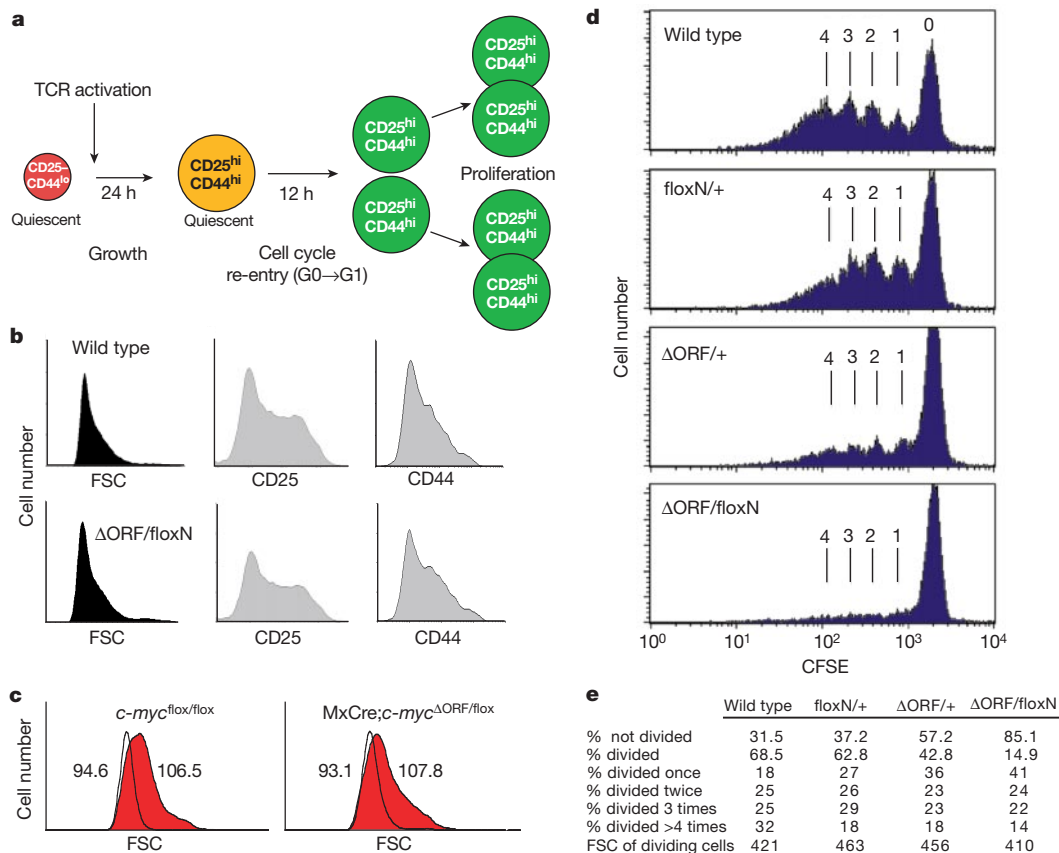


Figure 4 Activation, growth and proliferation potential of *c-myc* mutant naive CD4⁺ T cells. **a**, Diagram of the activation of peripheral T cells following stimulation of the T cell receptor (TCR). **b**, Size (FSC) and expression of CD25 and CD44 in T cells 24 h after TCR stimulation with anti-CD3 and anti-CD28. **c**, FSC after TCR stimulation with ConA of T cells isolated from *c-myc*^{flox/flox} mice and from MxCre; *c-myc*^{ΔORF/flox} mice in which Cre

expression was induced by polyI-polyC. Average FSC values of T cells before and after TCR activation are indicated for each genotype. **d**, *In vitro* proliferation assay (3 days) using TCR-stimulated CFSE-labelled naive CD4⁺ T cells isolated from littermates with the genotypes indicated. The number of divisions (0–4) are indicated above each graph. **e**, Quantification of the data shown in **d**.

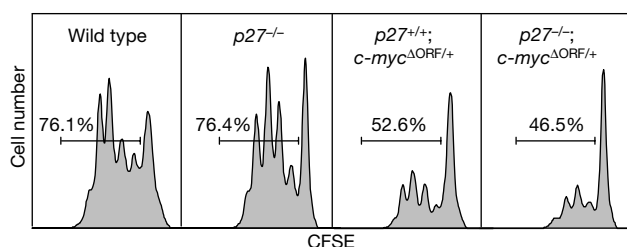


Figure 5 Proliferation potential of *c-myc*^{ΔORF/+} T cells in the absence of *p27*. *In vitro* proliferation assay using CFSE-labelled T cells isolated from wild-type and *c-myc*^{ΔORF/+} mice on a normal or *p27*^{-/-} background. The percentage of cells that re-entered the cell cycle in response to TCR activation is indicated above each graph.

growth phase. However, *c-Myc* levels determine the percentage of activated T cells that re-enter the cell cycle. In addition, deletion of *c-myc* from primary MEFs caused total arrest of cell proliferation. These data provide strong evidence that in mice, endogenous *c-Myc* functions as a switch that promotes entry into and prevents exit from the cell division cycle. Recent time-lapse video microscopy studies show that increased population doubling time of *c-myc*-deficient Rat1 cells¹⁵ is due to heterogeneity in the culture. Although a large proportion of daughter cells exit the cell cycle after mitosis, the cells that continue dividing have a doubling time similar to that of control cells²⁷. Furthermore, *c-myc* appears to be required for B cell proliferation and proper inactivation of Rb (ref. 6). Together those results support our conclusion that *c-Myc* in mammals is not essential for cellular growth and instead controls the decision of cells to divide or not to divide. Studies showing that ectopic expression of *c-Myc* alone is sufficient to promote cell-cycle re-entry of quiescent Rat1 fibroblasts in the absence of serum mitogens¹⁶, and that endogenous *c-Myc* expression is high during G0–G1 re-entry, low in proliferating cells and off in most resting cells¹⁷, are consistent with this hypothesis.

Although the data thus appear to suggest that *Myc* has different functions in *Drosophila* and mice, this is contradicted by evidence that despite their relatively weak sequence homology, both *c-Myc* and *dMyc* proteins have similar biological activities^{8,18}. Our results showing that *dmyc* expression can at least partially rescue the proliferation defect in *c-myc*-deficient mouse fibroblasts (Fig. 3d) support this view and further suggest that in mouse cells *dMyc* can control target genes normally regulated by *c-Myc*. The difference between flies and mice may therefore lie in the identity of the target genes that are controlled by *Myc* activity in each organism, or the way in which those target genes are integrated in the genetic circuitry of that organism. The opposite phenotypes, organ hypotrophy in *Drosophila* versus hypoplasia in mice, may therefore result from differences in the way invertebrates and mammals regulate tissue and body size. Among insect species, organ and body size differences appear to be a function of cell number and cell size, whereas among mammalian species they are almost exclusively due to variations in cell number^{19–21}. This may be due to a tighter coupling of cell growth and cell division in mammals than in insects²². Such a link would maintain average cell size in an expanding population and would thus make tissue size determination in mammals a function of the number of cell divisions and hence a function of *Myc* activity. □

Methods

Detailed methods are available in the Supplementary Information.

Generation and analysis of mutant mice

All *c-myc* alleles were generated by standard gene targeting in HM-1 ES cells²³. For inducible deletion of *c-myc*^{lox}, mice carrying MxCre were treated with poly(I)-poly(C) (Sigma)^{10,11}. Haematopoietic progenitor colony and RNase protection assays were essentially done as described^{24,25}.

Isolation and culture of primary mouse embryonic fibroblasts

Primary MEFs and 3T9 fibroblasts were isolated and cultured using standard protocols⁷. The retroviruses pMIG, pMIG-Cre⁺ and pBAGE-dmyc-puro were produced in BOSC 23 cells as previously described²⁶.

T-cell purification and *in vitro* proliferation assays

Naive T cells were purified from pooled spleen and lymph nodes collected from *c-myc* mutant mice, as described previously²⁶. TCR-induced proliferation was assayed by incubating 10⁶ CFSE-stained CD4⁺ T cells with 1 μg ml⁻¹ soluble anti-CD3 (clone 2C11, Pharmingen) and 10 μg ml⁻¹ soluble anti-CD28 (clone 37N1, Pharmingen), or alternatively with Concanavalin A (ConA) for three days, and then determining by flow cytometry the distribution of cells that have completed different numbers of cell divisions.

Received 4 September; accepted 19 October 2001.

- Grandori, C., Cowley, S. M., James, L. P. & Eisenman, R. N. The *Myc*/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* **16**, 653–699 (2000).
- Johnston, L. A., Prober, D. A., Edgar, B. A., Eisenman, R. N. & Gallant, P. *Drosophila myc* regulates cellular growth during development. *Cell* **98**, 779–790 (1999).
- Davis, A. C., Wims, M., Spotts, G. D., Hann, S. R. & Bradley, A. A null *c-myc* mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice. *Genes Dev.* **7**, 671–682 (1993).
- Orkin, S. H. & Zon, L. I. Genetics of erythropoiesis: induced mutations in mice and zebrafish. *Annu. Rev. Genet.* **31**, 33–60 (1997).
- Van Parijs, L., Refaelli, Y., Abbas, A. K. & Baltimore, D. Autoimmunity as a consequence of retrovirus-mediated expression of C-FLIP in lymphocytes. *Immunity* **11**, 763–770 (1999).
- de Alboran, I. M. *et al.* Analysis of *c-myc* function in normal cells via conditional gene-targeted mutation. *Immunity* **14**, 45–55 (2001).
- Todaro, G. J. & Green, H. Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J. Cell Biol.* **17**, 299–313 (1963).
- Gallant, P., Shilo, Y., Cheng, P. F., Parkhurst, S. M. & Eisenman, R. N. *Myc* and Max homologs in *Drosophila*. *Science* **274**, 1523–1527 (1996).
- Crabtree, G. R. Contingent genetic regulatory events in T lymphocyte activation. *Science* **243**, 355–361 (1989).
- Kuhn, R., Schwenk, F., Aguet, M. & Rajewsky, K. Inducible gene targeting in mice. *Science* **269**, 1427–1429 (1995).
- Radtke, F. *et al.* Deficient T cell fate specification in mice with an induced inactivation of Notch1. *Immunity* **10**, 547–558 (1999).
- Vlach, J., Hennecke, S., Alevizopoulos, K., Conti, D. & Amati, B. Growth arrest by the cyclin-dependent kinase inhibitor *p27^{Kip1}* is abrogated by *c-Myc*. *EMBO J.* **15**, 6595–6604 (1996).
- O'Hagan, R. C. *et al.* *Myc*-enhanced expression of *cull1* promotes ubiquitin-dependent proteolysis and cell cycle progression. *Genes Dev.* **14**, 2185–2191 (2000).
- Fero, M. L. *et al.* A syndrome of multiorgan hyperplasia with features of gigantism, tumorigenesis, and female sterility in *p27^{Kip1}*-deficient mice. *Cell* **85**, 733–744 (1996).
- Mateyak, M. K., Obaya, A. J., Adachi, S. & Sedivy, J. M. Phenotypes of *c-Myc*-deficient rat fibroblasts isolated by targeted homologous recombination. *Cell Growth Differ.* **8**, 1039–1048 (1997).
- Eilers, M., Schirm, S. & Bishop, J. M. The *myc* protein activates transcription of the α -prothymosin gene. *EMBO J.* **10**, 133–141 (1991).
- Blackwood, E. M., Kretzner, L. & Eisenman, R. N. *Myc* and Max function as a nucleoprotein complex. *Curr. Opin. Genet. Dev.* **2**, 227–235 (1992).
- Schreiber-Agus, N. *et al.* *Drosophila Myc* is oncogenic in mammalian cells and plays a role in the diminutive phenotype. *Proc. Natl Acad. Sci. USA* **94**, 1235–1240 (1997).
- Stevenson, R. D., Hill, M. F. & Bryant, P. J. Organ and cell allometry in Hawaiian *Drosophila*: how to make a big fly. *Proc. R. Soc. Lond. B* **259**, 105–110 (1995).
- Conlon, I. & Raff, M. Size control in animal development. *Cell* **96**, 235–244 (1999).
- Potter, C. J. & Xu, T. Mechanisms of size control. *Curr. Opin. Genet. Dev.* **11**, 279–286 (2001).
- Stocker, H. & Hafen, E. Genetic control of cell size. *Curr. Opin. Genet. Dev.* **10**, 529–535 (2000).
- Selfridge, J., Pow, A. M., McWhir, J., Magin, T. M. & Melton, D. W. Gene targeting using a mouse HPRT minigene/HPRT-deficient embryonic stem cell system: inactivation of the mouse ERCC-1 gene. *Somat. Cell Mol. Genet.* **18**, 325–336 (1992).
- Shivdasani, R. A., Mayer, E. L. & Orkin, S. H. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* **373**, 432–434 (1995).
- Trumpp, A., Blundell, P. A., de la Pompa, J. L. & Zeller, R. The chicken limb deformity gene encodes nuclear proteins expressed in specific cell types during morphogenesis. *Genes Dev.* **6**, 14–28 (1992).
- Refaelli, Y., Van Parijs, L., London, C. A., Tschopp, J. & Abbas, A. K. Biochemical mechanisms of IL-2-regulated Fas-mediated T cell apoptosis. *Immunity* **8**, 615–623 (1998).
- Hölzel, M., Kohlgruber, F., Schlosser, I., Hölzel, D., Lüscher, B. & Eick, E. *Myc*/Max/Mad regulate the frequency but not the duration of productive cell cycles. *EMBO Rep.*, 21 November 2001 (10.1093/embo-reports/kve251).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank J. Roberts and M. Fero for providing the *p27^{Kip1}* mice, S. Dymecki for the β -actin-FLP mice, D. Melton for HM1 ES cells, P. Gallant for *dmyc* cDNA, and L. VanParijs for pMIG-Cre. We thank G. Yee for performing the RNase protection analysis, S. Kogan for help with the methylcellulose cultures, D. Ginzinger for help with the initial Taqman analysis and C. McArthur, P. Zaech and A. Wilson for FACS sorting. We thank A. C. Pasche and D. Trail for technical assistance. We also thank M. Nabholz, S. Martin and our colleagues in the Trumpp, Bishop and Martin laboratories for discussions and critical reading of the manuscript. A.T. was the recipient of postdoctoral fellowships from the Deutsche Forschungsgemeinschaft, the Human Frontiers in Science Program, and the

California Division of the American Cancer Society. A.T. is now supported by grants from the Swiss National Science Foundation and the Swiss Cancer League. Y.R. is a Merck fellow of the Life Sciences Research Foundation. This work was supported by the NIH (G.R.M. and J.M.B.), and the G.W. Hooper Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.T. (e-mail: Andreas.Trump@isrec.unil.ch).

Bone indentation recovery time correlates with bond reforming time

James B. Thompson*, Johannes H. Kindt*, Barney Drake*, Helen G. Hansma*, Daniel E. Morse† & Paul K. Hansma*

* Department of Physics and † Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, California 93106, USA

Despite centuries of work, dating back to Galileo¹, the molecular basis of bone's toughness and strength remains largely a mystery. A great deal is known about bone microstructure^{2–5} and the microcracks^{6,7} that are precursors to its fracture, but little is known about the basic mechanism for dissipating the energy of an impact to keep the bone from fracturing. Bone is a nanocomposite of hydroxyapatite crystals and an organic matrix. Because rigid crystals such as the hydroxyapatite crystals cannot dissipate much energy, the organic matrix, which is mainly collagen, must be involved. A reduction in the number of collagen cross links has been associated with reduced bone strength^{8–10} and collagen is molecularly elongated ('pulled') when bovine tendon is strained¹¹.

Using an atomic force microscope^{12–16}, a molecular mechanistic origin for the remarkable toughness of another biocomposite material, abalone nacre, has been found¹². Here we report that bone, like abalone nacre, contains polymers with 'sacrificial bonds' that both protect the polymer backbone and dissipate energy. The time needed for these sacrificial bonds to reform after pulling correlates with the time needed for bone to recover its toughness as measured by atomic force microscope indentation testing. We suggest that the sacrificial bonds found within or between collagen molecules may be partially responsible for the toughness of bone.

Such sacrificial bonds, in abalone nacre, are analogous to those in the muscle protein titin^{13–16}. The energy needed to break a polymer designed in this way can be hundreds or even thousands of times greater than the energy needed to break a covalent bond, because the polymer must be re-stretched every time a sacrificial bond breaks and releases more 'hidden length'^{13–16}. The hidden length can be, as in titin^{13–16}, the difference in length between an unfolded and a folded domain along one polymer chain or the extra length released when sacrificial bonds between two chains break.

Force versus extension curves—'pulls'—of collagen show that there are sacrificial bonds that can be broken, and thereby prevent the force from rising to a value that would break the collagen backbone (Fig. 1). If, after pulling on collagen, the tip of the atomic force microscope (AFM) is moved back to within approximately 50 nm of the surface, sacrificial bonds will reform. The longer the delay before the next pull, the more sacrificial bonds reform and thus the more energy dissipation observed when pulling the collagen molecule(s) again. If the collagen is in a calcium buffer, more sacrificial bonds reform and more energy dissipation is observed than if it is in a sodium buffer. Even so, less than half of the energy dissipation observed the first time collagen is pulled is regained by allowing a 100-s delay in calcium buffer. Energy dissipation also tends to decrease as collagen is repeatedly pulled, suggesting that either some of the sacrificial bonds cease to reform, or that one or more of several molecules attached to the tip become

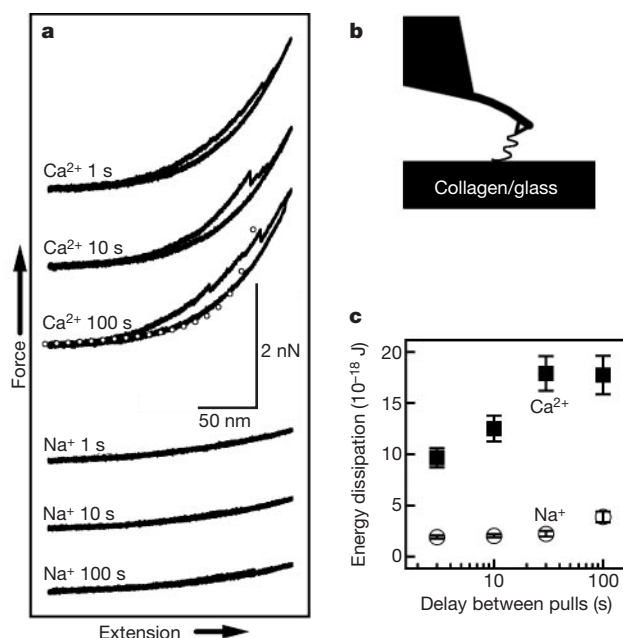


Figure 1 Pulling on collagen. **a**, Force versus extension curves obtained by pulling collagen molecules supported on a glass cover slip. The example curves shown were selected from a series of pulls on the same molecule(s) to have average values of energy dissipation, that is average areas enclosed by the extension away from the surface (upper trace) and return to the surface (lower trace). This energy dissipation changes with the delay between successive pulls, noted on the curves, on the same molecule(s). For

calcium buffer, the area increases with delay. For sodium buffer, the area is smaller and increases less with delay. The return portion of the Ca²⁺ 100-s curve was fitted to the wormlike-chain model (open circles). The model does not fit well, suggesting that we are not pulling on a single protein backbone. **b**, Diagram of the experiment. **c**, A plot of the average energy dissipation versus delay time. Each curve, Ca²⁺ and Na⁺, was calculated from the enclosed area of several pulls on the same molecule(s).

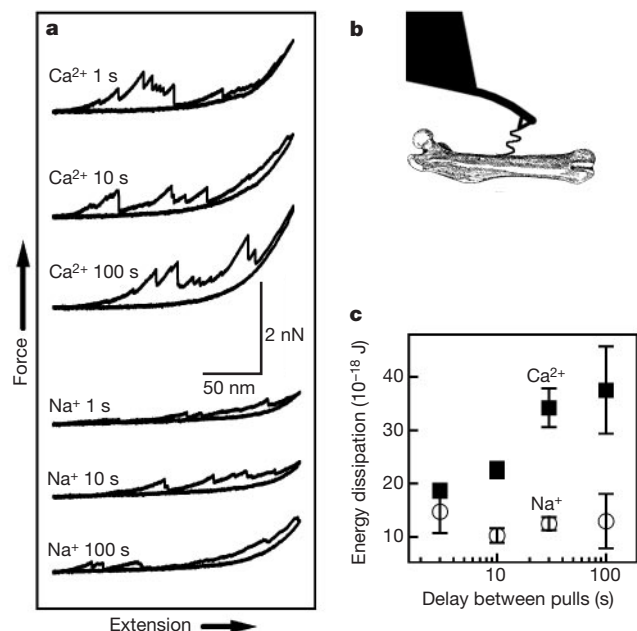


Figure 2 Pulling on bone. **a**, Force versus extension curves measured on a polished piece of fresh rat femur. **b**, Diagram of the experiment. **c**, As for Fig. 1c, measured on collagen, the energy dissipation measured on rat femur in a calcium buffer increases with the delay between pulls. We note that the timescale for this increase as a function of the delay between pulls is comparable to that in Fig. 1b.

detached. We do not know the nature of these sacrificial bonds. One possibility is that they include ionic bridges between two negatively charged ions on the collagen such as carboxylate ions. Divalent calcium ions can form such bridges, but monovalent sodium ions cannot.

Our collagen pulling data did not readily fit the wormlike-chain model^{13,17} typically applied to pulling curves (Fig. 1). The fits implied persistence lengths of only 0.03 nm. Such a persistence length is unphysical and less than a tenth of the value typically quoted for pulling on single proteins^{13–15}. Either we are pulling on several molecules in parallel, or collagen has unusual force versus extension properties. The collagen backbone is folded into a triple helix, which might not become unfolded during extension. Therefore, when pulling collagen, we might be stretching three interacting protein chains, even when pulling on a single molecule. If so, the wormlike chain would be an inappropriate model for pulling curves on collagen.

This research on pulling collagen complements existing research on collagen with AFMs. After early pioneering papers^{18,19} there have been over thirty papers reporting the use of AFM to image collagen from sources as diverse as tendon²⁰, dentin²¹ and cornea^{22,23}. Both assembly^{22,24} and degradation have been studied^{25,26}. Pulling curves on collagen have also been reported²⁷, but the ability of collagen to recover energy dissipation with time was not investigated.

It is not necessary to use purified collagen to observe the behaviour shown in Fig. 1. Similar behaviour is seen when pulling molecules exposed on a polished bone (rat femur) surface (Fig. 2). At this time, we cannot prove that these pulling curves are due to pulling of collagen. However, pulling curves for polished bone are similar to those for purified collagen and exhibit a similar recovery of energy dissipation as a function of the delay between pulls, suggesting that we are pulling on similar molecules in both cases. Further, collagen is the most abundant polymer found in bone and we find the behaviour shown in Fig. 2 to be ubiquitous on the surface of bone.

We wondered whether this mechanism of energy dissipation has any relationship to energy dissipation or strength in bone. Our first experiment in this area involved soaking bovine femur samples with dimensions of order 2 cm × 2 mm × 2 mm in either 1 M NaCl or

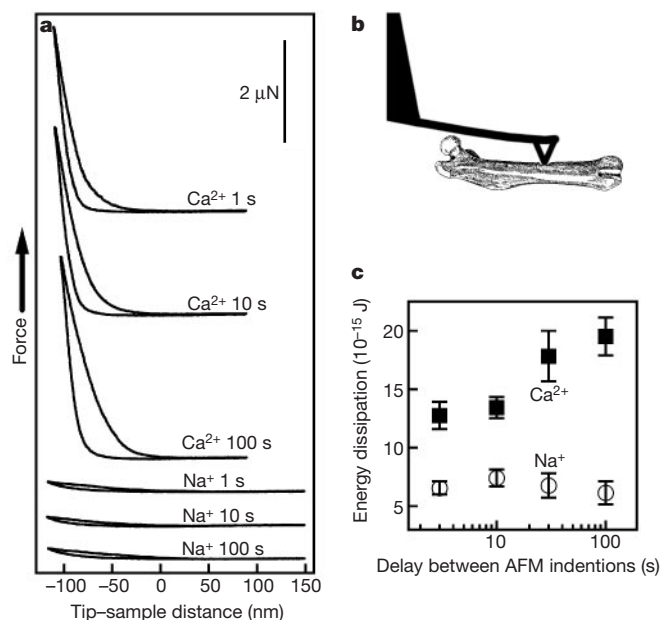


Figure 3 Indentation of bone. **a**, Non-destructive indentation curves on fresh rat femur probed at forces low enough that no permanent deformation of the surface occurs. In calcium buffer the bone is stronger, that is, the force versus tip sample distance curves are steeper. As in Figs 1 and 2, the area under the curves—energy dissipation—increases with increasing delay between indentations. **b**, Diagram of the experiment. **c**, Average energy dissipation versus delay time. The energy dissipation tends to be larger in calcium buffer and increases with a similar timescale to that observed in Figs 1 and 2. One major difference is that here the energy dissipation is roughly three orders of magnitude greater than in the molecular pulling curves even though the indentation distances are smaller than the pull distances. If the same molecules are involved, this implies that many more molecules are acting together to resist indentation than we ‘pulled’ in the pulling experiment.

1 M CaCl_2 for several days and then performing four point bending tests. These tests showed that the yield strength of samples soaked in 1 M CaCl_2 was not significantly different than the yield strength of samples soaked in 1 M NaCl. Thus, we were led to the conclusion that either this mechanism of energy dissipation had nothing to do with the yield strength of bone, or that the ions in our two solutions had not completely penetrated our samples and exchanged with the ions already present.

In an attempt to determine which was the correct answer, we tried a more local probe of the strength and toughness of bone. We used a very stiff AFM cantilever (50 N m^{-1}) to indent rat femur samples, and found significant differences between samples soaked in calcium buffer and those soaked in sodium buffer (Fig. 3). On average, samples soaked in calcium buffer appeared harder, and when they were indented by about 50 nm these samples recovered their initial energy dissipation with a similar timescale to that observed in collagen. It is important to note that indenting the bone surface by about 50 nm did not produce a permanent deformation, which suggests that the bone’s structure was not greatly altered. When the surface was indented by about 100 nm, a permanent deformation of the surface occurred, and energy dissipation did not recover within 100 s. AFM indentation is only sensitive to changes of bone properties near the surface, yet we found it necessary to soak bone samples for a day or more before they reached steady-state properties. It appears that the exchange of ions is slow within bone, which supports the hypothesis that our macroscopic bone tests showed no differences owing to insufficient penetration and exchange of the solution ions with those buried in the bone. It has been found that AFM indentation of primary osteoblasts results in increased intracellular calcium concentration²⁸.

For sacrificial bonds in collagen to dissipate energy during

indentation, indentation must result in sufficient elongation of collagen to break these sacrificial bonds. We do not yet understand how this occurs. Collagen within bone is assembled into fibrils²⁻⁴ and thereby held in an extended configuration. We speculate that collagen in this extended configuration may be pre-stressed so that sacrificial bonds must be broken in order to allow even small indentations in the bone surface.

It seems surprising that bonds may form at the same rate within collagen supported on a glass slide and collagen in bone. Perhaps the formation of sacrificial bonds requires only the presence of multivalent ions, which bind two sites on the same or (more probably) two different collagen molecules. Solution ions were given several days to penetrate the bone samples, so the concentration of multivalent ions within the outer layers of the bone samples should be similar to that seen by collagen molecules exposed directly to the solution.

Thus, we have shown that in bone that has been indented, in molecules from bone that have been pulled and in purified collagen that has been pulled, the recovery of toughness takes about the same amount of time. This correlation in time dependence suggests, but does not prove, that these recoveries may be related. The recovery of toughness in our pulling experiments is in some ways similar to that seen in experiments on abalone shell proteins¹² and titin¹³⁻¹⁶. Although there are almost certainly not any precisely folded domains as in titin, there do appear to be renewable sacrificial bonds that break at forces below the force needed to break the backbone of these polymers. These sacrificial bonds provide a mechanism for toughness; it takes much more energy to pull these molecules than it would in the absence of sacrificial bonds. This property of collagen is of interest because collagen is the most abundant protein in the human body and serves as a structural component of a variety of tissues including bone, tendon, and skin.

Note added in proof: Initial experiments indicate that pulling curves taken on collagen molecules in PBS also show a recovery of energy dissipation with time delay between pulls (our own unpublished data). Thus, multivalent negative ions, such as phosphate, also appear to form sacrificial bonds. □

Methods

AFM pulling

The pulling experiments were conducted using a prototype small-cantilever AFM. The cantilevers used typically had spring constants of 50 pN nm⁻¹, resonant frequencies of 120 kHz and 0.5-μm-long electron-beam-deposited tips. Coating the cantilever tip with gold did not affect the results reported. The pulling curves shown were taken with a 30-kHz bandwidth and at a loading rate of 2 μm s⁻¹. Our typical root mean square (r.m.s.) noise level was less than 0.3 pN Hz^{-1/2}. The force was measured by multiplying the deflection of the cantilever by the spring constant of the cantilever. Extension is measured from the starting point of the pulls: approximately 50 nm off the surfaces. During each pull the molecules were extended by 200 nm. Pulling curves were captured in a series with 1-s delays between pulls of: 1, 1, 100, 1, 30, 1, 10, 1, 3, 1, 3, 1, 10, 1, 30, 1, 100, 1, 100, 1, 30, 1, 10, 1, 3, 1, 3, 1, 10, 1, 30, 1 and 100 seconds (in that order), without contacting the surface. The tip was kept approximately 50 nm away from the surface during the entire series to avoid picking up additional molecules. Energy dissipation was calculated and averaged for all curves in a single series to produce the data sets shown in Figs 1b and 2b.

AFM indentation

The AFM indentation tests were conducted using a commercial atomic force microscope (Nanoscope III MultiMode from Digital Instruments) and cantilever (Nanosensors). These curves were taken over a 2-s period. The surface was typically raised 50 nm beyond the point where the cantilever first made contact, but it was necessary to reduce this distance on particularly hard samples. Force was calculated by multiplying cantilever deflection by the spring constant of the cantilever, in this case approximately 50 N m⁻¹. Tip sample distance refers to the motion of the sample, which is raised for the indentation, minus the deflection of the cantilever. Positive tip-sample distances correspond to the tip being above the surface of the sample. Negative tip-sample distances correspond to the tip indenting below the original surface of the sample ('zero'). Indentation curves were captured in a series with delays between indentations of either: 1, 1, 100, 1, 30, 1, 10 and 3 s or the reverse order (increasing from 3 to 100 s). Energy dissipation was calculated and averaged for curves taken at five or more positions, separated from each other by 20 μm, on the bone surface to produce the data sets shown in Fig. 3b.

Buffers

Each measurement was conducted in one of two buffers. The calcium buffer contained

40 mM CaCl₂, 110 mM NaCl, and 10 mM HEPES. The sodium buffer contained 150 mM NaCl and 10 mM HEPES. The pH of both solutions was adjusted to 7.0 using a small amount of NaOH.

Collagen sample preparation

Collagen samples were prepared by suspending ~1 mg of acid insoluble collagen from bovine Achilles tendon (Sigma)²⁹ in a 10 μl drop of PBS (phosphate buffered saline) on a cover glass. A second cover glass was used to shear the collagen on the first cover glass. The PBS was then allowed to evaporate to dryness. The dried sample was rinsed under a stream of deionized water (Milli-Q) for 15 s before placing either sodium or calcium buffer on the sample and introducing it into the AFM.

Bone sample preparation

The bone samples used in the pulling and indentation experiments were prepared from the mid-section of a rat femur. The rat had been killed less than 1 h before sample preparation. The femur was first cleaned and cut into pieces of about 2 mm × 2 mm × 0.5 mm. The 2 mm × 2 mm surface was sanded with 400 grit and then 600 grit silicon carbide sandpaper. After sanding, the surface was polished with 30 μm, 6 μm, 1 μm and finally 0.25 μm diamond suspensions before thoroughly rinsing it in Milli-Q water. This left a clean, flat surface of longitudinally cut compact bone. The samples were stored under buffer. The pulling curves shown were taken on the first and second day after sample preparation. The samples were soaked in the buffer solutions at 6 °C for 5 days before AFM indentation testing.

Received 29 August; accepted 4 October 2001.

1. Asencio, A. Biomechanics and Galileo Galilei. *J. Biomech.* **26**, 95–100 (1993).
2. Weiner, S. & Wagner, H. D. The material bone: structure-mechanical function relations. *Annu. Rev. Mater. Sci.* **28**, 271–298 (1998).
3. Weiner, S., Traub, W. & Wagner, H. D. Lamellar bone: structure-function relations. *J. Struct. Biol.* **126**, 241–255 (1999).
4. An, Y. H. & Draughn, R. A. *Mechanical Testing of Bone and the Bone-Implant Interface* (CRC Press, New York, 2000).
5. Currey, J. D. *The Mechanical Adaptations of Bones* (Princeton Univ. Press, Princeton, 1984).
6. Reilly, G. C. & Currey, J. D. The effects of damage and microcracking on the impact strength of bone. *J. Biomech.* **33**, 337–343 (2000).
7. Reilly, G. C. & Currey, J. D. The development of microcracking and failure in bone depends on the loading mode to which it is adapted. *J. Exp. Biol.* **202**, 543–552 (1999).
8. Oxlund, H., Barckman, M., Ortoft, G. & Andreassen, T. T. Reduced concentrations of collagen cross-links are associated with reduced strength of bone. *Bone* **17**, S365–S371 (1995).
9. Burstein, A. H., Zika, J. M., Heiple, K. G. & Klein, L. Contribution of collagen and mineral to the elastic-plastic properties of bone. *J. Bone Joint Surg. A* **57**, 956–961 (1975).
10. Knott, L. & Bailey, A. J. Collagen cross-links in mineralizing tissues: A review of their chemistry, function, and clinical relevance. *Bone* **22**, 181–187 (1998).
11. Sasaki, N. & Odajima, S. Elongation mechanism of collagen fibrils and force-strain relations at each level of structural hierarchy. *J. Biomech.* **29**, 1131–1136 (1996).
12. Smith, B. L. *et al.* Molecular mechanistic origin of toughness of natural adhesives, fibre and composites. *Nature* **399**, 761–763 (1999).
13. Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J. M. & Gaub, H. E. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* **276**, 1109–1112 (1997).
14. Rief, M., Gautel, M., Schemmel, A. & Gaub, H. E. The mechanical stability of immunoglobulin and fibronectin II domains in the muscle protein titin measured by atomic force microscopy. *Biophys. J.* **75**, 3008–3014 (1998).
15. Li, H., Oberhauser, A. F., Fowler, S. B., Clarke, J. & Fernandez, J. M. Atomic force microscopy reveals the mechanical design of a modular protein. *Proc. Natl Acad. Sci. USA* **97**, 6527–6531 (2000).
16. Oberhauser, A. F., Hansma, P. K., Carrion-Vazquez, M. & Fernandez, J. M. Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proc. Natl Acad. Sci. USA* **98**, 468–472 (2001).
17. Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of lambda-phage DNA. *Science* **265**, 1599–1600 (1994).
18. Chernoff, E. A. G. & Chernoff, D. A. Atomic force microscope images of collagen fibers. *J. Vac. Sci. Technol. A* **10**, 596–599 (1994).
19. Revenko, I., Sommer, F., Minh, D. T., Garrone, R. & Franc, J. M. Atomic force microscopy study of the collagen fiber structure. *Biol. Cell* **80**, 67–69 (1994).
20. Bigi, A., Gandolfi, M., Roveri, N. & Valdre, G. *In vitro* calcified tendon collagen: an atomic force and scanning electron microscopy investigation. *Biomaterials* **18**, 657–665 (1997).
21. El Feninat, F., Ellis, T. H., Sacher, E. & Stangel, I. Moisture-dependent renaturation of collagen in phosphoric acid etched human dentin. *J. Biomed. Mater. Res.* **42**, 549–553 (1998).
22. Miyagawa, A. *et al.* Surface topology of collagen fibrils associated with proteoglycans in mouse cornea and sclera. *Jpn J. Ophthalmol.* **44**, 591–595 (2000).
23. Fullwood, N. J., Hammiche, A., Pollock, H. M., Hourston, D. J. & Song, M. Atomic force microscopy of the cornea and sclera. *Curr. Eye Res.* **14**, 529–535 (1995).
24. Paige, M. F. & Goh, M. C. Ultrastructure and assembly of segmental long spacing collagen studied by atomic force microscopy. *Micron* **32**, 355–361 (2001).
25. Lin, H., Clegg, D. O. & Lal, R. Imaging real-time proteolysis of single collagen I molecules with an atomic force microscope. *Biochemistry* **38**, 9956–9963 (1999).
26. Yamamoto, S. *et al.* The subfibrillar arrangement of corneal and scleral collagen fibrils as revealed by scanning electron and atomic force microscopy. *Arch. Histol. Cytol.* **63**, 127–135 (2000).
27. Dufrene, Y. F., Marchal, T. G. & Rouxhet, P. G. Influence of substratum surface properties of the organization of adsorbed collagen films: *in situ* characterization by atomic force microscopy. *Langmuir* **15**, 2871–2878 (1999).
28. Charras, G. T., Lehenkari, P. P. & Horton, M. A. Atomic force microscopy can be used to mechanically stimulate osteoblasts and evaluate cellular strain distributions. *Ultramicroscopy* **86**, 86–95 (2001).
29. Einbinder, J. & Schubert, M. Binding of mucopolysaccharides and dyes by collagen. *J. Biol. Chem.* **188**, 335–341 (1951).
30. Greene, E. C. *Anatomy of the Rat* (Hafner, New York, 1963).

Acknowledgements

We thank E. Oroudjev, J. Cooper, D. Kohn and L. Fisher for their assistance and discussion. We also thank the reviewers of our paper for their helpful suggestions. We thank S. Babbitt and the American Philosophical Society for granting us permission to use the drawing of a rat femur³⁰ shown in Figs 2 and 3. This work was supported by the Materials Research Division and MCB Division of the National Science Foundation, the Materials Research Laboratory Program of the National Science Foundation, and the MURI programme of the Army Research Office.

Correspondence and requests for materials should be addressed to J.B.T. (e-mail: jbtthomp@physics.ucsb.edu).

Bacteriophytochromes are photochromic histidine kinases using a biliverdin chromophore

Seong-Hee Bhoo[†]*, Seth J. Davis[†]*, Joseph Walker^{*}, Baruch Karniol^{*} & Richard D. Vierstra^{*}

^{*} Cellular and Molecular Biology Program and the Department of Horticulture, University of Wisconsin—Madison, 1575 Linden Drive, Madison, Wisconsin 53706, USA

[†] These authors contributed equally to this work

Phytochromes comprise a principal family of red/far-red light sensors in plants¹. Although phytochromes were thought originally to be confined to photosynthetic organisms^{2,3}, we have recently detected phytochrome-like proteins in two heterotrophic eubacteria, *Deinococcus radiodurans* and *Pseudomonas aeruginosa*⁴. Here we show that these form part of a widespread family of bacteriophytochromes (BphPs) with homology to two-component sensor histidine kinases. Whereas plant phytochromes use phytychromobilin as the chromophore, BphPs assemble with biliverdin, an immediate breakdown product of haem, to generate photochromic kinases that are modulated by red and far-red light. In some cases, a unique haem oxygenase responsible for the synthesis of biliverdin is part of the BphP operon. Co-expression of this oxygenase with a BphP apoprotein and a haem source is sufficient to assemble holo-BphP *in vivo*. Both their presence in many diverse bacteria and their simplified assembly with biliverdin suggest that BphPs are the progenitors of phytochrome-type photoreceptors.

Photosynthetic organisms use an array of photoreceptor systems to optimize their growth and development to the ambient light environment. The phytochromes of plants form a family of homodimeric chromoproteins, which contain the linear tetrapyrrole 3E-phytychromobilin (PΦB) covalently attached to polypeptides of about 1,100 amino acids¹. These photoreceptors sense red (R) and far-red (FR) light through photo-interconversion between two stable conformations, an R-absorbing Pr form, and an FR-absorbing Pfr form. Because Pfr is biologically active but Pr is inactive, phytochromes act as reversible R/FR 'switches' that synchronize many aspects of photomorphogenesis¹.

We have identified strong sequence homology between the previously reported phytochrome-like proteins from *D. radiodurans* and *P. aeruginosa*⁴ and the *BphP* loci from the Gram-negative bacteria *Pseudomonas syringae* strain DC3000, *Pseudomonas putida*, *Pseudomonas fluorescens*, *Agrobacterium tumefaciens* and *Rhizobium leguminosarium*, the α -proteobacterium *Rhodospirillum rubrum*, the previously described phytochrome-like photoreceptors from the α -proteobacterium *Rhodospirillum rubrum*⁵ and several cyanobacteria^{6,7}, and BphP-like sequences from the fungi *Neurospora crassa* and *Aspergillus fumigatus* (Fig. 1a). Like phytochromes, these bacteriophytochrome photoreceptors (BphPs) contain an amino-terminal chromophore-binding domain, which autocatalytically attaches bilins, followed by a histidine kinase module^{3,4}. BphPs are notably distinct from phytochromes as they lack the cysteine used by phytochromes to bind bilins through a thioether bond¹. Instead, BphPs seem to use the adjacent conserved histidine to bind the bilin through a Schiff-base-type linkage (ref. 4; and Fig. 1a). This protein organization suggested that BphPs, like cyanobacterial phytochromes^{8,9}, function as light-activated kinases similar to other bacterial two-component sensors¹⁰.

Whereas higher plant and cyanobacterial phytochromes are known to use respectively PΦB and 3Z-phytycyanobilin (PCB) as the chromophore^{1,9}, the chromophore of BphP was unclear because bacteria other than cyanobacteria and fungi are not known to produce such linear bilins¹¹. To determine whether the BphP chromophore is a bilin, we isolated the native BphP photoreceptor from wild-type *D. radiodurans* and assayed it for this type of chromophore by zinc-induced fluorescence¹². To enrich for *Dr* BphP, we used anti-*Dr* BphP antibodies to concentrate it directly from a crude soluble lysate. The enriched protein fluoresced in the presence of zinc and ultraviolet light similar to holo-BphP

from wild-type *D. radiodurans* and assayed it for this type of chromophore by zinc-induced fluorescence¹². To enrich for *Dr* BphP, we used anti-*Dr* BphP antibodies to concentrate it directly from a crude soluble lysate. The enriched protein fluoresced in the presence of zinc and ultraviolet light similar to holo-BphP

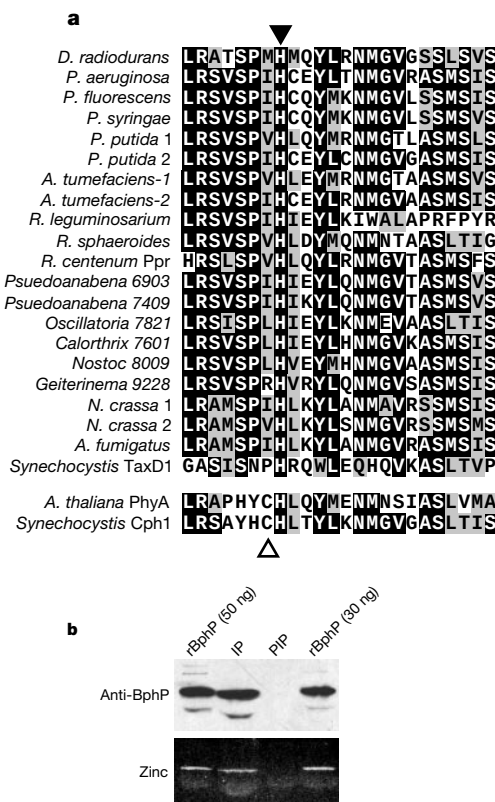


Figure 1 Sequence comparison of the BphP family and detection of the BphP chromoprotein from *D. radiodurans*. **a**, Amino-acid sequence alignment of the region surrounding the predicted chromophore-binding site for BphPs and phytochromes. Filled and open arrowhead identifies the histidine and cysteine residue involved in binding the bilin in BphPs and the phytochromes (*A. thaliana* phyA and *Synechocystis* Cph1), respectively. Reverse type and grey boxes denote identical and similar amino acids, respectively. **b**, Detection of BphP from *D. radiodurans*. *Dr* BphP was immuno-affinity purified from crude cell extracts with either anti-*Dr* BphP antibodies (IP) or pre-immune serum (PIP) and subject to SDS-PAGE. The polypeptide was detected by immunoblot analysis with anti-*Dr* BphP antibodies, and the bound bilin chromophore was detected by zinc-induced fluorescence. *Dr* BphP bound to BV *in vitro* was used as a standard.

[†] Present address: Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK.

synthesized *in vitro*, indicating that the native *Dr* BphP contains a covalently bound bilin (Fig. 1b).

In higher plants and cyanobacteria, P Φ B and PCB are synthesized from haem by multi-step pathways involving the oxidation of haem by a haem oxygenase to form biliverdin IX α (BV), followed by reduction and isomerization steps¹³. Extensive searches using as queries cyanobacterial and plant sequences for haem oxygenase and bilin reductases^{13,14} failed, however, to identify possible orthologues in any of the eubacterial species that contain BphPs. With the expectation that haem is the precursor for the BphP chromophore, these eubacteria should contain an activity that cleaves this cyclic tetrapyrrole to generate BV. One possibility was HemO, an enzyme related to but substantially distinct from haem oxygenases, which is used by *Neisseriae gonorrhoeae* to acquire iron by catabolizing haemin¹⁵. Using HemO or its *P. aeruginosa* orthologue, PigA, as queries^{15,16}, we identified related proteins (designated BphO for bacteriophytochrome haem oxygenase) in eubacteria that also contain a BphP (Fig. 2b; and Supplementary Information). Unexpectedly, in several species the *BphO* gene immediately precedes the *BphP* gene in an operon (Fig. 2a), suggesting that there is a functional linkage between haem metabolism and photoreceptor function.

Often the *BphP* operons include additional open-reading frames that might encode immediate downstream component(s) of this

sensory cascade (Fig. 2a; and data not shown). Like the *Synechocystis* *Phy* operon encoding *Cph1* (ref. 9), the *D. radiodurans*, *P. putida* and *R. leguminosarum* *BphP* operons encode a response regulator (designated BphR) related to *Escherichia coli* CheY that presumably functions as the phosphoacceptor for the activated kinase⁴.

Both higher plant and cyanobacterial phytochromes are unable to bind BV (ref. 17; and Fig. 2c), an attachment precluded by the A-ring vinyl side chain. But given the presence of closely linked *BphO* genes and the use of a Schiff's-base-type linkage, which makes attachment of BV more chemically feasible, we reasoned that BphPs might incorporate this bilin. To examine this possibility, we tested several recombinant BphPs for their ability to attach BV autocatalytically. As expected, the representative phytochrome *Synechocystis* *Cph1* did not bind BV, but did bind PCB and 3Z-phycoerythrobilin (PEB) (Fig. 2c). Under the same conditions, the BphP apoproteins from *D. radiodurans*, *P. syringae* and *P. aeruginosa* covalently bound all three bilins; in fact, BV was coupled with an efficiency equal to or better than that of PCB and PEB (Fig. 2c; and data not shown).

As predicted⁴, neither the *D. radiodurans* nor the *P. syringae* BphPs bound BV when an alanine was substituted for the positionally conserved histidine (see Fig. 1a), indicating that BV probably becomes attached to this residue (data not shown). The BphPs generated a non-photochromic product when assembled with PEB, but generated repeatedly R/FR photochromic products when assembled with BV and PCB (data not shown). Whereas the PCB-containing BphPs had difference spectra uncharacteristic of phytochromes, including a minimum in the green region of the spectrum and substantial variation for the maximum and minimum in the red region of the spectrum, the spectra of the BV-containing BphP photoreceptors were identical to each other and similar in shape to phytochromes assembled with PCB or P Φ B^{1,9}. This suggested that BV 'fits better' into the chromophore-binding

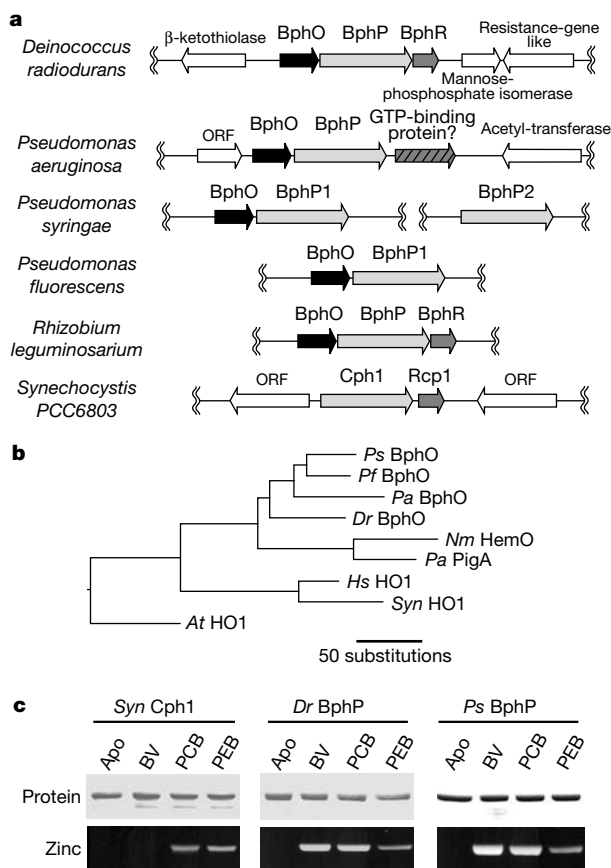


Figure 2 Detection of sequences encoding haem oxygenase (BphOs) in several *BphP* operons and assembly of BphPs with BV. **a**, Organization of several *BphP* operons containing *BphO*, and comparison with the *Cph1* operon from *Synechocystis*.

b, Phylogenetic comparison of BphOs from *P. syringae* (Ps), *P. fluorescens* (Pf), *P. aeruginosa* (Pa) and *D. radiodurans* (Dr) with HemO from *N. gonorrhoeae*, PigA from *P. aeruginosa*, and with haem oxygenases from *Synechocystis* (Syn), *Arabidopsis thaliana* (At) and *Homo sapiens* (Hs). **c**, *In vitro* assembly of BphPs with various bilins. Recombinant versions of *D. radiodurans* and *P. syringae* BphPs and *Synechocystis* *Cph1* were incubated with BV, PCB or PEB. The reaction products were subjected to SDS-PAGE, and either stained for protein with Coomassie blue or assayed for the bound bilin by zinc-induced fluorescence.

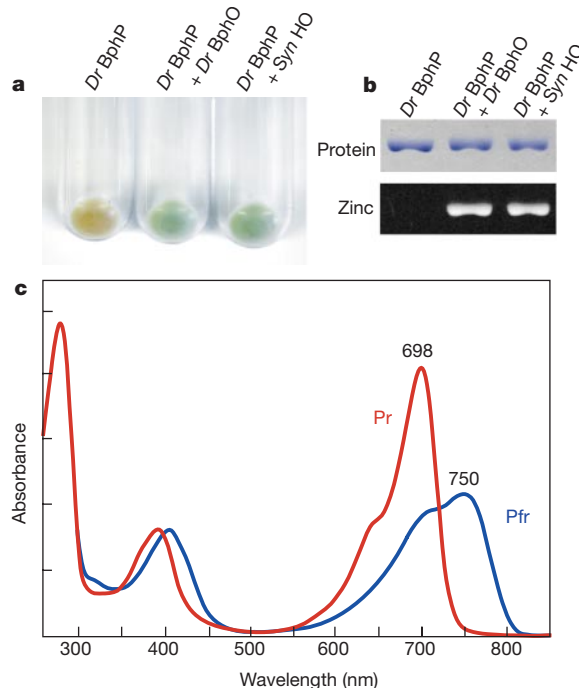


Figure 3 *In vivo* assembly of recombinant *D. radiodurans* BphP by co-expression of the apoprotein with *D. radiodurans* (Dr) BphO or *Synechocystis* (Syn) haem oxygenase 1.

a, Packed *E. coli* cells transformed with the designated plasmids, showing the blue-green colour of the accumulating *Dr* BphP holoprotein after induction. **b**, Isolation of the recombinant *Dr* BphP purified from the induced bacteria shown in **a**. The *Dr* BphP protein was subjected to SDS-PAGE, and either stained for protein by Coomassie blue (top) or analysed for the bound bilin by zinc-induced fluorescence (bottom). **c**, Absorbance spectra after saturating R and FR of *Dr* BphP holoprotein assembled with BV *in vivo*.

pocket of BphPs than do other bilins.

To prove that BphOs synthesize BV, we developed a coupled assay that assembles holo-BphP in *E. coli* by co-expressing *Dr* BphP with *Dr* BphO. *E. coli* cells expressing *Dr* BphP alone were the same pale yellow colour as wild-type cells. When expression of *Dr* BphP was combined with either *Dr* BphO or a *bona fide* haem oxygenase from *Synechocystis* (*Syn* HO)¹⁸, however, the cells became bright blue-green, the expected colour of the BV-BphP holoprotein, and contained a *Dr* BphP protein bearing a covalently attached bilin (Fig. 3a, b). Purified *Dr* holo-BphP displayed Pr and Pfr absorbance spectra identical to that of *Dr* BphP assembled *in vitro* with BV, confirming that BV is the bound chromophore, and by inference, that *Dr* BphO is a genuine haem oxygenase (Fig. 3c). These BV-BphP spectra were remarkably similar to those for phytochromes assembled with P Φ B except for a significant red-shift of the Pr and Pfr absorbance peaks (Fig. 3c).

Assembly of holo-BphP in the presence of either oxygenase appeared to be saturated for BV, as we were unable to convert additional apoprotein to a spectrally active chromoprotein, either by adding BV to BphP purified from these cells or by supplementing the culture medium with BV before induction. Yields of about 20 mg of *Dr* holo-BphP per litre of *E. coli* culture were easily obtained, indicating that this recombinant system may be particularly useful for studies that require ample quantities of homogeneous holoproteins.

We confirmed the possibility that BphPs are light-regulated histidine kinases by autophosphorylation assays using recombinant *Ps* BphP assembled with BV. The apoprotein had low intrinsic kinase activity, indicating that assembly with bilin is necessary to activate the histidine kinase domain. Whereas phytochromes from higher plants and cyanobacteria appear to be more active kinases as the Pr form^{9,19}, kinetic analysis of *Ps* BphP indicated the opposite:

autophosphorylation was roughly 20 times faster for Pfr than for Pr (Fig. 4a, b). The kinase activity was photoreversible by R/FR/R cycles, a property characteristic of phytochromes, and was more active when assembled with BV than with PCB or PEB (Fig. 4a).

Consistent with His532 serving as the phosphoacceptor site, the autophosphorylated form of *Ps* BphP was base stable and acid labile (Fig. 4c). Because recombinant *Ps* BphP behaves as a homodimer (data not shown), we assume on the basis of other sensor histidine kinases¹⁰ that phosphorylation proceeds in a *trans* process between the subunits. Once phosphorylated, *Ps* BphP then could donate the phosphate to a response regulator, in this case *Dr* BphR (Fig. 4d). Phosphotransfer to *Dr* BphR was slightly more efficient as Pfr than as Pr, once the autophosphorylated intermediate was established as Pfr.

Taken together, we conclude that BphPs represent a new class of light-activated histidine kinases related to phytochromes that use BV as the chromophore. The use of BV is supported by (1) the association of a haem oxygenase responsible for BV synthesis in many *BphP* operons; (2) the ability of representative BphPs to bind BV *in vitro* or assemble with BV generated *in vivo* by co-expression with a haem oxygenase; (3) the similar spectral properties of BphP chromoproteins containing BV and those of phytochromes; and (4) the ability of BV-BphP to function as a light-regulated kinase. Whereas the more complex bilins such as PCB and P Φ B have well-described roles in phytochromes and photosynthetic accessory pigments^{1,11}, this represents, to our knowledge, the first documented role for BV in photoperception.

The resulting BV-BphP chromoproteins have R/FR photochromic absorbance spectra similar to those of phytochromes, indicating strong evolutionary conservation of their spectral characteristics despite using a different chromophore bound by a different linkage. In contrast to phytochromes^{9,19}, the one BphP

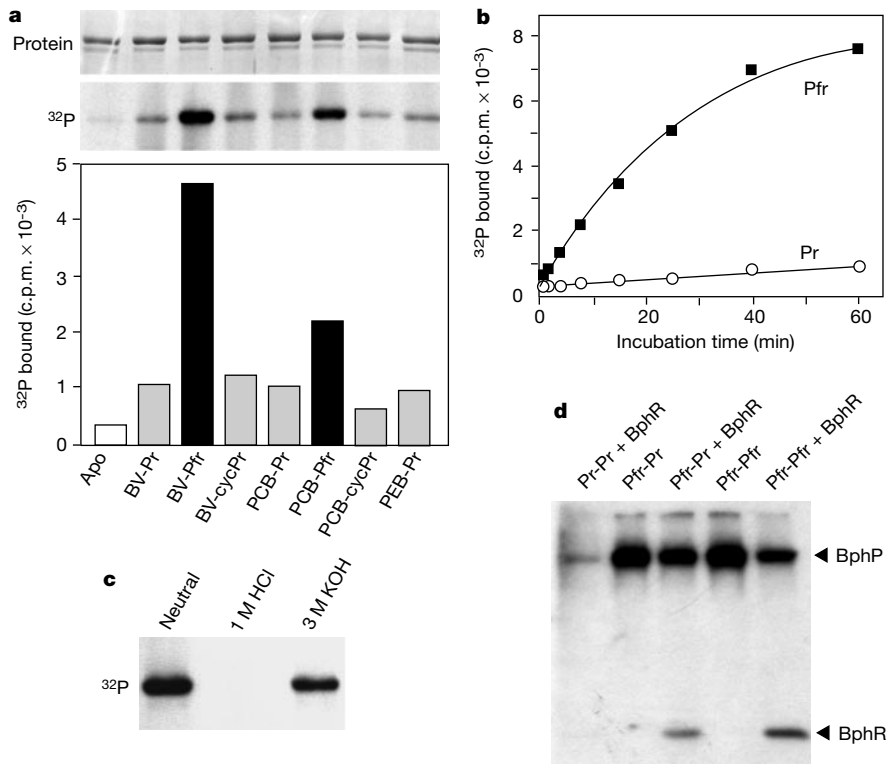


Figure 4 *Ps* BphP assembled with BV acts as a R-activated histidine kinase in a two-component phospho-relay system. **a**, Autophosphorylation of purified BphP without chromophore (Apo) or assembled with BV, PCB or PEB and photo-converted to Pr or Pfr. After incubation with [γ -³²P]ATP, the reaction products were subjected to SDS-PAGE, and stained with Coomassie blue (top) and subjected to autoradiography (middle). Bound ³²P was measured by scintillation counting of the *Ps* BphP protein excised from the gel

(bottom). **b**, Autophosphorylation kinetics of the Pr and Pfr forms of *Ps* BphP assembled with BV. **c**, Stability of the autophosphorylated form of *Ps* BphP in 1 M HCl or 3 M KOH. **d**, Transfer of the phosphate from *Ps* BphP to *Dr* BphR. *Ps* BphP assembled with BV was autophosphorylated for 40 min as Pr or Pfr. BphP was either left in those forms (Pr-Pr, Pfr-Pfr) or the Pfr was cycled back to Pr (Pfr-Pr), and then incubated with 10 μ M unlabelled ATP and *Dr* BphR for an additional 5 min. The products were visualized as in **a**.

tested here (*P. syringae* BphP) was a more active kinase as the Pfr form. Although the reason for this difference is unknown, the greater kinase activity for the Pfr form of this BphP conforms with the long-held expectation from physiological studies that Pfr is the biologically active form of phytochrome-type photoreceptors¹.

Since our detection of BphPs in *D. radiodurans* and *P. aeruginosa*⁴, we have now discovered related sequences in several heterotrophic and photoautotrophic bacteria and in fungi, indicating that this family of photoreceptors is part of a widely used photosensory system. Given this distribution, we predict that BphPs represent the progenitor of phytochromes. Why was BV replaced by PCB and PΦB during the ontogeny of phytochromes? One reason may be reflected by the substantial R shift in the Pr absorbance maximum of the BV–BphP chromoproteins as compared with the Pr absorbance maximum of phytochromes. Whereas phytochromes are perfectly adapted for measuring shade avoidance by photosynthetic organisms because their Pr spectrum closely overlaps that of chlorophylls¹, BV–BphPs are not because their Pr maxima lie at 700 nm. As a consequence, cyanobacteria and plants might have adopted blue-shifted bilins such as PCB and PΦB to detect competition better, and thus enhance photosynthetic light capture. This switch would require an ability to make and bind these bilins, and an ability to prevent coupling of their precursor BV. Such a conversion may have been facilitated in cyanobacteria by their exploitation of PCB in photosynthetic light harvesting¹¹. □

Methods

Isolation and detection of BphP from *D. radiodurans*

Cells were grown as described⁴, collected by centrifugation, frozen and lysed in 100 mM Tris-HCl (pH 7.0), 140 mM ammonium sulphate, 10 mM EDTA, 50% ethylene glycol, 2 mM phenylmethylsulphonylfluoride. The clarified crude extract was incubated at 4 °C with anti-*Dr* BphP antibodies coupled to Affigel 10 (Pharmacia). Bound protein was eluted with 100 mM glycine (pH 2.5). The eluate was precipitated at –20 °C with eight volumes of acetone plus 0.002% 2-mercaptoethanol, collected by centrifugation, and resuspended in SDS–PAGE sample buffer. We performed immunoblot analysis with Hybond-C Extra membranes (Amersham) blocked with 1×PBS and 10% dried milk, and used horseradish peroxidase linked to goat anti-rabbit immunoglobulins (Kiergaard and Perry) for detection. Covalently bound bilins were visualized by zinc-induced fluorescence of the complexes subjected to 10% SDS–PAGE¹².

Expression and assembly of BphPs

BphP-coding regions were amplified by polymerase chain reaction from genomic DNA using primers designed to add appropriate restriction sites for expression as His₆-tagged versions in suitable pET expression vectors (Novagen). With the exception of *Dr* BphP, the His₆ domain was appended to the carboxy terminus. For *Dr* BphP, *Nde*I sites were added for insertion into pET28a. For *Ps* BphP, *Pa* BphP, *Dr* BphO, *Syn* HO and *Dr* BphR, *Bam*HI and *Xho*I sites were added for insertion into pET21a. The coding region for the His₆ tag was subsequently removed from each haem oxygenase by appropriate digestion.

Recombinant proteins were expressed at 22 °C in the dark in the *E. coli* strain BL21-CodonPlus (DE3)-RIL (Stratagene) upon induction with 1 mM isopropyl β-D-thiogalactoside. Co-expression of *Dr* BphP with either *Dr* BphO or *Syn* HO was achieved by co-transformation of appropriate plasmids and selection on LB medium containing 50 μg ml^{–1} ampicillin and 30 μg ml^{–1} of kanamycin. We grew the cultures at 37 °C until a roughly 0.4 absorbance at 600 nm was reached, and then induced and incubated them at 22 °C for 4 h. Cells were lysed in 30 mM Tris-HCl (pH 7.8), 300 mM NaCl, 20 mM imidazole. The BphPs were purified from the clarified crude extract by nickel/nitrilotriacetic acid affinity chromatography (Qiagen) using 30 mM Tris-HCl (pH 7.8), 300 mM NaCl, 200 mM imidazole for elution. The eluants were exchanged into 70 mM Tris-HCl (pH 8.0), 1 mM EDTA, 200 mM NaCl, 5% ethylene glycol. For BphPs assembled *in vitro*, the purified apoproteins were incubated for 1 h at 4 °C in a 10-fold molar excess of BV (Porphyrin Products), PCB or PEB (provided by P.-S. Song). Absorbance and difference spectra of BphPs were determined after saturating irradiations with light at 690 and 760 nm. We performed protein kinase reactions as described⁹. Sequences were compared using MACBOXSHADE (Institute of Animal Health, Pirbright, UK) and ClustalW (<http://www.ebi.ac.uk/clustalw>).

Received 18 July; accepted 15 October 2001.

- Smith, H. Phytochromes and light signal perception by plants—an emerging synthesis. *Nature* **407**, 585–591 (2000).
- Hughes, J. & Lamparter, T. Prokaryotes and phytochrome. The connection to chromophores and signaling. *Plant Physiol.* **121**, 1059–1068 (1999).
- Vierstra, R. D. & Davis, S. J. Bacteriophytochromes: new tools for understanding phytochrome signal transduction. *Semin. Cell. Dev. Biol.* **11**, 511–521 (2000).
- Davis, S. J., Vener, A. V. & Vierstra, R. D. Bacteriophytochromes: phytochrome-like photoreceptors from nonphotosynthetic eubacteria. *Science* **286**, 2517–2520 (1999).

- Jiang, Z. *et al.* Bacterial photoreceptor with similarity to photoactive yellow protein and plant phytochromes. *Science* **285**, 406–409 (1999).
- Herdman, M., Coursin, T., Rippka, R., Houmard, J. & Tandeau de Marsac, N. A new appraisal of the prokaryotic origin of eukaryotic phytochromes. *J. Mol. Evol.* **51**, 205–213 (2000).
- Bhaya, D., Takahashi, A. & Grossman, A. R. Light regulation of type IV pilus-dependent motility by chemosensor-like elements in *Synechocystis* PCC6803. *Proc. Natl Acad. Sci. USA* **98**, 7540–7545 (2001).
- Kehoe, D. M. & Grossman, A. R. Similarity of a chromatic adaptation sensor to phytochrome and ethylene receptors. *Science* **273**, 1409–1412 (1996).
- Yeh, K. C., Wu, S. H., Murphy, J. T. & Lagarias, J. C. A cyanobacterial phytochrome two-component light sensory system. *Science* **277**, 1505–1508 (1997).
- West, A. H. & Stock, A. M. Histidine kinases and response regulator proteins in two-component signaling systems. *Trends Biochem. Sci.* **26**, 369–376 (2001).
- Beale, S. I. Biosynthesis of phycobilins. *Chem. Rev.* **93**, 785–802 (1993).
- Berkelman, T. R. & Lagarias, J. C. Visualization of bilin-linked peptides and proteins in polyacrylamide gels. *Anal. Biochem.* **156**, 194–201 (1986).
- Davis, S. J., Kurepa, J. & Vierstra, R. D. The *Arabidopsis thaliana* HY1 locus, required for phytochrome-chromophore biosynthesis, encodes a protein related to heme oxygenases. *Proc. Natl Acad. Sci. USA* **96**, 6541–6546 (1999).
- Frankenberg, N., Mukougawa, K., Kohchi, T. & Lagarias, J. C. Functional genomic analysis of the h2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. *Plant Cell* **13**, 965–978 (2001).
- Zhu, W., Hunt, D. J., Richardson, A. R. & Stojiljkovic, I. Use of heme compounds as iron sources by pathogenic *Neisseria* requires the product of the *hemO* gene. *J. Bacteriol.* **182**, 439–447 (2000).
- Ochsner, U. A. & Vasil, M. L. Gene repression by the ferric uptake regulator in *Pseudomonas aeruginosa*: cycle selection of iron-regulated genes. *Proc. Natl Acad. Sci. USA* **93**, 4409–4414 (1996).
- Elich, T. D. & Lagarias, J. C. Formation of a photoreversible phycocyanobilin–apophytochrome adduct *in vitro*. *J. Biol. Chem.* **264**, 12902–12908 (1989).
- Cornejo, J., Willows, R. D. & Beale, S. I. Phytyl-bilin biosynthesis: cloning and expression of a gene encoding soluble ferredoxin-dependent heme oxygenase from *Synechocystis* sp. PCC 6803. *Plant J.* **15**, 99–107 (1998).
- Yeh, K. C. & Lagarias, J. C. Eukaryotic phytochromes: light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc. Natl Acad. Sci. USA* **95**, 13976–13981 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank P.-S. Song for providing PCB and PEB, and M. Wexler, J. Todd and A. Johnston for making the *R. leguminosarium* sequences available before publication. This work was supported by grants from the US Department of Energy and the National Science Foundation to R.D.V.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.D.V. (e-mail: vierstra@facstaff.wisc.edu). GenBank accession codes for the BphP sequences are AF418994 (*P. fluorescens*), AF418997 (*P. syringae*), AF418995 and AF418996 (*P. putida*), AAK87748 and AAK87910 (*A. tumefaciens*), AJ416905 (*R. leguminosarium*) and AF418998 (*R. sphaeroides*). Incomplete genome sequences are available for *N. crassa* (<http://www.genome.wi.mit.edu/annotation/fungi/neurospora>) and *A. fumigatus* (<http://www.tigr.org/tbd/e2k1/afu1>).

erratum

The rhythm of microbial adaptation

Philip Gerrish

Nature **413**, 299–302 (2001).

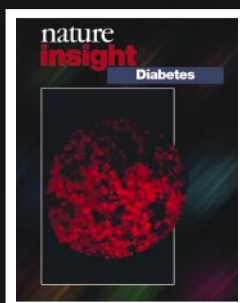
The final paragraph of the Methods was inadvertently omitted.

Analysis of HIV sequences

Eighteen sequences from the c2v3 region of the *env* gene of HIV-1 subtype E were derived from viral samples taken from different infected individuals in Thailand at approximately the same time²⁶. Codon alignment and phylogenetic analysis were performed using CLUSTAL²⁷ and PHYLIP²⁸. Twelve of the sequences formed a star phylogeny, suggesting independent evolution, and the founding sequence was determined. Non-synonymous distances were calculated with SNAP²⁹. The good agreement between this data and the prediction of equation (2) suggests that non-synonymous evolution in this gene region is mostly adaptive, an observation supported by many other studies of this region. □

nature insight

Diabetes



Cover illustration

An islet of Langerhans isolated from the pancreas and stained for insulin producing cells. (Image courtesy of C. Rhodes, Pacific Northwest Research Institute.)

Familiarity breeds contempt. This was true for the fox in Aesop's fable whose respect and fear for the mighty lion was dulled by seeing him day after day. Arguably the same has occurred with our attitude to diabetes. We hope that this collection of reviews by leading researchers in diabetes will help rectify this situation.

Diabetes is so prevalent that all readers of this Insight will know several diabetics, at least one of whom will be injecting insulin to keep their condition under control. At base, diabetes is caused by failure to maintain a stable level of blood glucose in the face of the normal fluctuations of supply and demand. However, the causes of this relatively straightforward dysfunction are anything but simple, and diabetes is not one but a spectrum of disorders.

The following articles discuss both type 1 diabetes, in which the insulin producing cells in the pancreas are killed early in life, and the more insidious type 2 diabetes, where daily environmental and dietary stresses overload and ultimately destroy the body's systems for regulating blood glucose. Both these forms, if uncontrolled, can lead to severe complications such as heart disease, kidney failure, blindness and even loss of limbs. The debate over the biochemical causes of these complications and the best ways to prevent their development is also reflected in these pages.

Despite a wealth of knowledge about diabetes we must not be complacent. The World Health Organization predicts a doubling in its incidence in the next two decades, fuelled predominantly by modern lifestyles and an increasing incidence of obesity. It is also acquiring a variety of new names, including metabolic syndrome, syndrome X, diabetes and insulin resistance. But whatever it is called diabetes is a complex and potentially debilitating condition which, like Aesop's lion, must be taken very seriously.

We are pleased to acknowledge the financial support of Bristol-Myers Squibb in producing this Insight. As always, *Nature* retains sole responsibility for editorial content and its peer review.

Christopher Surridge Senior Editor
Deepa Nath Associate Editor

review articles:

- 782** **Global and societal implications of the diabetes epidemic**
*P. Zimmet,
K. G. M. M. Alberti
& J. Shaw*

- 788** **Diabetes mellitus and genetically programmed defects in β -cell function**
G. I. Bell & K. S. Polonsky

- 792** **β -Cell death during progression to diabetes**
*D. Mathis, L. Vence
& C. Benoist*

- 799** **Insulin signalling and the regulation of glucose and lipid metabolism**
A. R. Saltiel & C. R. Kahn

- 807** **Mitochondrial function in normal and diabetic β -cells**
*P. Maechler
& C. B. Wollheim*

- 813** **Biochemistry and molecular cell biology of diabetic complications**
M. Brownlee

- 821** **New drug targets for type 2 diabetes and the metabolic syndrome**
D. E. Moller

Editor, *Nature*: Philip Campbell
Insight Editors: Christopher Surridge
Deepa Nath
Lesley Anson

Editorial Assistant: Simon Gibson
Art Director: Majo Xeridat
Diagrams: Ann Thomson
Suzanne Coleman

Production Editor: Simon Gribbin
Production Manager: Yvonne Strong
Sponsorship: Emma Jones

Global and societal implications of the diabetes epidemic

Paul Zimmet*, K. G. M. M. Alberti† & Jonathan Shaw*

*International Diabetes Institute, 260 Kooyong Road, Caulfield, Victoria 3162, Australia (e-mail: pzimmet@idi.org.au)

†Royal College of Physicians, 11 St Andrews Place, Regent's Park, London NW1 4LE, UK

Changes in human behaviour and lifestyle over the last century have resulted in a dramatic increase in the incidence of diabetes worldwide. The epidemic is chiefly of type 2 diabetes and also the associated conditions known as 'diabesity' and 'metabolic syndrome'. In conjunction with genetic susceptibility, particularly in certain ethnic groups, type 2 diabetes is brought on by environmental and behavioural factors such as a sedentary lifestyle, overly rich nutrition and obesity. The prevention of diabetes and control of its micro- and macrovascular complications will require an integrated, international approach if we are to see significant reduction in the huge premature morbidity and mortality it causes.

"Man may be the captain of his fate, but he is also the victim of his blood sugar"

Wilfrid Oakley [*Trans. Med. Soc. Lond.* **78**, 16 (1962)]

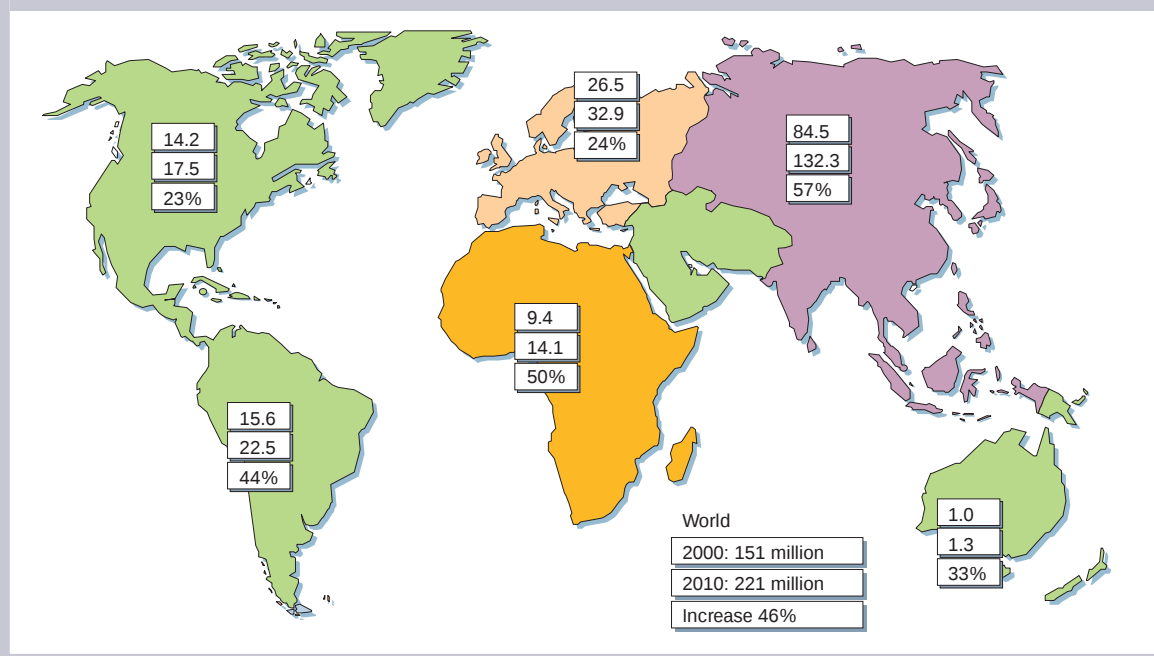
Diabetes mellitus, long considered a disease of minor significance to world health, is now taking its place as one of the main threats to human health in the 21st century¹. The past two decades have seen an explosive increase in the number of people diagnosed with diabetes worldwide^{2,3}. Pronounced changes in the human environment, and in human behaviour and lifestyle, have accompanied globalization, and these have resulted in escalating rates of both obesity and diabetes. Hence the recent adoption of the term 'diabesity'⁴, first suggested by Shafir several decades ago⁵.

There are two main forms of diabetes⁶. Type 1 diabetes is due primarily to autoimmune-mediated destruction of

pancreatic β -cell islets, resulting in absolute insulin deficiency. People with type 1 diabetes must take exogenous insulin for survival to prevent the development of ketoacidosis. Its frequency is low relative to type 2 diabetes, which accounts for over 90% of cases globally. Type 2 diabetes is characterized by insulin resistance and/or abnormal insulin secretion, either of which may predominate. People with type 2 diabetes are not dependent on exogenous insulin, but may require it for control of blood glucose levels if this is not achieved with diet alone or with oral hypoglycaemic agents.

The diabetes epidemic relates particularly to type 2 diabetes, and is taking place both in developed and developing nations⁷. Paradoxically, part of the problem relates to the achievements in public health during the 20th century, with

Figure 1 Numbers of people with diabetes (in millions) for 2000 and 2010 (top and middle values, respectively), and the percentage increase. Data adapted from ref. 2.



people living longer owing to elimination of many of the communicable diseases⁸. Non-communicable diseases (NCD) such as diabetes and cardiovascular disease (CVD) have now become the main public health challenge for the 21st century, as a result of their impact on personal and national health and the premature morbidity and mortality associated with the NCDs¹.

After taking so long to gain recognition, interest in diabetes is now mounting rapidly⁷ and it is an exciting time for researchers and clinicians involved in the study and treatment of the disease. The problem has crept up on an unsuspecting public health community. The global figure of people with diabetes is set to rise from the current estimate of 150 million to 220 million in 2010, and 300 million in 2025 (Fig. 1)^{2,3}. Most cases will be of type 2 diabetes, which is strongly associated with a sedentary lifestyle and obesity⁷. This trend of increasing prevalence of diabetes and obesity has already imposed a huge burden on health-care systems and this will continue to increase in the future^{1,9}.

Although type 2 diabetes is numerically more prevalent in the general population, type 1 diabetes is the most common chronic disease of children. But with the increasing prevalence of type 2 diabetes in children and adolescents, the order may be reversed within one to two decades^{10,11}.

Prevention of complications

Prevention of complications is a key issue because of the huge premature morbidity and mortality associated with the disease^{1,12}. In the past decade, several major studies have focused attention on the need for strict control of glycaemia to prevent and/or reduce the risk of both the specific microvascular and the less specific macrovascular complications⁷.

The Diabetes Control and Complications Trial¹³ was a landmark study and the flagship for a number of studies that established the value of intensive control of blood glucose to prevent the retinal, renal and neuropathic complications of diabetes. The United Kingdom Prospective Diabetes Study (UKPDS)¹⁴ fulfilled the same role for type 2 diabetes. Subsequently, there were other important studies that underline the importance of active medical intervention (including control of blood pressure and lipids as well as glucose) for the reduction of the risk of diabetes complications. This applies to microvascular complications, as shown, for example, by the Stockholm¹⁵, MICRO-HOPE¹⁶ and Kumamoto studies¹⁷, and to macrovascular disease from the 4S¹⁸, CARE¹⁹ and MICRO-HOPE¹⁶ studies.

The concealed burden of impaired glucose tolerance

Type 2 diabetes is increasingly common, indeed epidemic, primarily because of increases in the prevalence of a sedentary lifestyle and obesity²⁰. The possibility of preventing type 2 diabetes by interventions that affect the lifestyles of subjects at high risk for the disease is now the subject of a number of studies; these have focused on people with impaired glucose tolerance (IGT)^{21–23}. IGT is defined as hyperglycaemia (with glucose values intermediate between normal and diabetes) following a glucose load (ref. 6 and Table 1), and affects at least 200 million people worldwide. It represents a key stage in the natural history of type 2 diabetes as these people are at much higher future risk than the general population for developing diabetes^{24,25}. Approximately 40% of subjects progress to diabetes over 5–10 years, but some revert to normal or remain IGT.

Subjects with IGT also have a heightened risk of macrovascular disease^{25–29}. Because of this, and the association with other known CVD risk factors including hypertension, dyslipidaemia and central obesity^{6,30}, the diagnosis of IGT, particularly in apparently healthy and ambulatory individuals, has important prognostic implications^{29,31}.

Impaired fasting glucose (IFG) was introduced recently as another category of abnormal glucose metabolism^{6,32}. It is defined on the basis of fasting glucose concentration (Table 1) and, like IGT, it is associated with risk of CVD and future diabetes. The American Diabetes Association had hoped that their recommendation of the

Table 1 Values for diagnosis* of diabetes and other types of hyperglycaemia

	Glucose concentration (mmol l ⁻¹)			
	Plasma Venous	Capillary	Whole blood Venous	Capillary
Diabetes mellitus				
Fasting	≥ 7.0	≥ 7.0	≥ 6.1	≥ 6.1
2-h post-glucose load	≥ 11.1	≥ 12.2	≥ 10.0	≥ 11.1
Impaired glucose tolerance				
Fasting concentration	< 7.0	< 7.0	< 6.1	< 6.1
2-h post-glucose load	7.8–11.0	8.9–12.1	6.7–9.9	7.8–11.0
Impaired fasting glucose				
Fasting	6.1–6.9	6.1–6.9	5.6–6.0	5.6–6.0
2-h post-glucose load	< 7.8	< 8.9	< 6.7	< 7.8

*Note that diabetes can be diagnosed in an individual only when these diagnostic values are confirmed on another day. Data from ref. 24.

Ranges of values are inclusive (that is, 6.1–6.9 means ≥ 6.1 and < 7.0).

adoption of IFG³² would simplify the screening for at-risk individuals, with the need for an oral glucose-tolerance test being circumvented. However, the desired outcome has yet to be achieved as IGT is a better predictor than IFG of risk of future diabetes and of mortality^{22,27,28,33}.

Type 2 diabetes and the metabolic syndrome

Type 2 diabetes is a multifactorial disease that shows heterogeneity in many respects³⁴. Our understanding of the disease and related disorders such as IGT and IFG is undergoing a radical change, particularly as data suggest that the risk of complications commences many years before the onset of clinical diabetes^{30,35}. Previously, it was regarded as a relatively distinct disease entity, but in reality, type 2 diabetes (and its associated hyperglycaemia or dysglycaemia) is often a manifestation of a much broader underlying disorder^{7,36}. This includes the metabolic syndrome (sometimes called syndrome X; Fig. 2) — a cluster of CVD risk factors that, in addition to glucose intolerance (that is, IGT or diabetes), includes hyperinsulinaemia, dyslipidaemia, hypertension, visceral obesity, hypercoagulability and microalbuminuria. This new paradigm relating to type 2 diabetes also influences contemporary therapy for the disease⁷. Evidence now exists for a far more aggressive approach to treating not just the hyperglycaemia, but also other CVD risk factors such as hypertension, dyslipidaemia and central obesity in type 2 diabetic

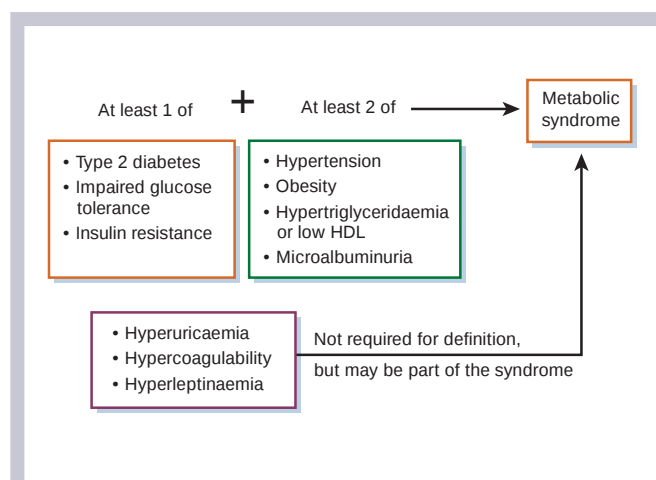


Figure 2 Metabolic syndrome as defined by the World Health Organization²⁴. Insulin resistance is defined as being within the highest quartile for the relevant population. Hypertension is defined as blood pressure ≥ 140/90. Obesity is defined as a body-mass index ≥ 30 kg m⁻², or a waist–hip ratio (WHR) > 0.90 for males and WHR > 0.85 for females. Hypertriglyceridaemia is defined as ≥ 1.7 mmol triglycerides l⁻¹. Low high-density lipoprotein (HDL) is defined as < 0.9 mmol l⁻¹ for men; < 1.0 mmol l⁻¹ for women. Microalbuminuria is a urinary albumin excretion rate ≥ 20 µg min⁻¹ or albumin creatinine ratio ≥ 30 mg min⁻¹.

patients, with the hope of significantly reducing cardiovascular morbidity and mortality.

There is continuing debate as to the primary aetiological factor for the syndrome. Genetic factors, visceral obesity, insulin resistance and endothelial dysfunction may all contribute either solely or jointly^{36,37}. Irrespective of the aetiological debate, a recent report from the WHO⁶ has highlighted both the need and the importance of a consistent definition of the metabolic syndrome and has suggested parameters. Although the criteria may change as new prospective data become available, this initiative provides a basis for the development of a standardized definition that will allow international comparisons of prevalence, incidence and natural history, as previously there was no internationally agreed definition.

The reliability of this definition in terms of predicting the prevalence of, and the CVD risk associated with, the metabolic syndrome has been reported recently³⁸. Cardiovascular mortality was assessed in 3,606 subjects from the Botnia study (a large-scale study of type 2 diabetes begun in Finland in 1990) with a median follow-up of 6.9 years. In women and men, respectively, the metabolic syndrome was recorded in 10 and 15% of subjects with normal glucose tolerance, 42 and 64% of those with IFG/IGT, and 78 and 84% of those with type 2 diabetes. The risk for coronary heart disease and stroke was increased threefold in subjects with the syndrome, and cardiovascular mortality was markedly increased (12.0% in subjects with the syndrome versus 2.2% in those without; $P < 0.001$). This study confirms the hope of the WHO report⁶ that the WHO definition of the metabolic syndrome will identify subjects with increased CVD morbidity and mortality. It offers a tool for comparison of results from different studies.

Diabetes in children and youth

Type 2 diabetes in children, teenagers and adolescents is a serious new aspect to the epidemic and an emerging public health problem of significant proportions^{10,11,39}. Although type 1 diabetes remains the main form of the disease in children worldwide, it seems possible that type 2 diabetes will be the predominant form within ten years in many ethnic groups and potentially in Europid groups (that is, of European descent). Type 2 diabetes has already been reported in children from Japan, the United States, Pacific Islands, Hong Kong, Australia and the United Kingdom^{10,39–41}. Among children in Japan, it is already more common than type 1 diabetes, accounting for 80% of childhood diabetes; the incidence almost doubled between 1976–80 and 1991–5 (ref. 42).

The rising prevalence of obesity and type 2 diabetes in children is symptomatic of the effects of globalization and industrialization affecting all societies¹, with sedentary lifestyle and obesity¹¹ the predominant factors involved. As a result of this new and alarming scenario, a joint consensus statement has been issued recently by the American Diabetes Association and the American Academy of Pediatrics⁴⁰. We now urgently need epidemiological data to outline the extent of the problem, as most of the reports so far are clinic based^{11,39}.

Although type 2 diabetes in Europeans is usually characterized by onset (often asymptomatic) after 50 years of age, in Pacific islanders and other high-prevalence groups such as southern Asians, onset in the 20–30-year age group is now increasingly common^{7,39}. The socioeconomic and public health impact of this downward shift in disease onset on society is much greater through effects on the work force and premature morbidity and mortality, as well as through the negative impact on fertility and reproduction⁴³.

The appearance of type 2 diabetes in a younger age group also raises new issues in classification of diabetes. For example, how do we differentiate type 2 diabetes in children from type 1? Certainly, the presence of C-peptide and absence of markers of autoimmunity such as antibodies to glutamic acid decarboxylase may help to diagnose type 2 diabetes⁴⁴. What therapies are safe in this age group? Most pharmacological therapies for diabetes and its associated conditions (apart from insulin) are not approved for use in children and the

Table 2 Aetiological determinants and risk factors of type 2 diabetes

Genetic factors
Genetic markers, family history, 'thrifty gene(s)'
Demographic characteristics
Sex, age, ethnicity
Behavioural- and lifestyle-related risk factors
Obesity (including distribution of obesity and duration)
Physical inactivity
Diet
Stress
'Westernization, urbanization, modernization'
Metabolic determinants and intermediate risk categories of type 2 diabetes
Impaired glucose tolerance
Insulin resistance
Pregnancy-related determinants (parity, gestational diabetes, diabetes in offspring of women with diabetes during pregnancy, intra-uterine mal- or overnutrition)

same applies for those for blood pressure and dyslipidaemia^{11,40}. How will we tackle the problem of diabetes complications at an earlier age? All of these issues need to be addressed urgently, and information on behavioural and environmental factors will be required to plan intervention strategies.

The main factors in the type 2 diabetes epidemic

The diabetes epidemic, although apparent right around the world, has been most pronounced in non-Europid populations, as evidenced by studies from Native American and Canadian communities, Pacific and Indian Ocean island populations⁴⁵, groups in India⁴⁶ and Australian Aboriginal communities⁴⁷. In the Pacific island of Nauru, where diabetes was virtually unknown 50 years ago, it is now present in approximately 40% of adults⁴⁸. The potential for increases in the number of cases of diabetes is greatest in Asia². Data from Mauritius show the highest yet reported prevalence in people of Chinese extraction, in addition to demonstrating a high diabetes prevalence and a notable secular increase between 1987 and 1998 in Asian Indians and Creoles⁷. Together with evidence that prevalence of type 2 diabetes doubled between 1984 and 1992 in Singaporean Chinese⁴⁹, and with the high prevalence in Taiwan⁵⁰, these data provide alarming indicators of the size of the future epidemic in the People's Republic of China. Here, the overall prevalence of type 2 diabetes was, until recently, less than 1%. Recent studies show a threefold increase in prevalence in certain areas of China within the past two decades⁵¹. The scenario is highlighted in Fig. 3 where the prevalence

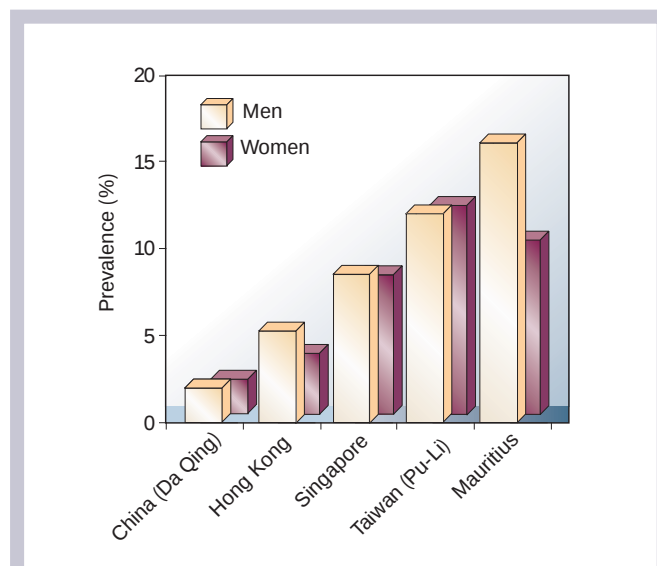


Figure 3 The prevalence of type 2 diabetes mellitus among Chinese in Hong Kong, Singapore, Taiwan and Mauritius, compared with that in the People's Republic of China⁶.

of diabetes in Chinese populations in China, Singapore, Taiwan and Mauritius is compared. If China were to experience just one-half of the current rate of diabetes in Taiwan, the number of individuals with diabetes will increase from 8 million in 1996 to over 32 million by 2010. The data from the Indian subcontinent are equally disconcerting^{7,46}.

What are the reasons for this epidemic? Apart from the heightened genetic susceptibility of certain ethnic groups, environmental and behavioural factors such as sedentary lifestyle, nutrition and obesity (Table 2) are clearly important⁵². One of the major debates in diabetes and NCD aetiology is the issue of thrifty genotype versus thrifty phenotype⁷. In one corner are those who postulate that humans are genetically programmed for the hunter-gatherer era. The 'thrifty genotype' hypothesis⁵³ provides an explanation of the very high prevalence of obesity and type 2 diabetes in the American Pima Indians, Australian Aborigines and Pacific Islanders — the basis for the susceptibility to diabetes could be the result of an evolutionary advantageous thrifty genotype that promoted fat deposition and storage of calories in times of plenty. This mechanism would have conferred a survival advantage during the regular famines and natural disasters that were interspersed with feast periods. With Westernization, these populations now have a plentiful supply of a diet with an excess of energy, simple carbohydrates and saturated fats. This has been accompanied by a reduction in both occupational and leisure-based physical activity. Both factors may therefore cause the previously favourable metabolic profile seen in 'survivors' to become a handicap, which results in obesity and type 2 diabetes⁵⁴.

There exists an excellent model of this phenomenon — the Israeli sand rat (*Psammomys obesus*)⁷. When this animal is removed from the sparse diet of its natural environment and given an abundant, high-calorie diet, it develops all of the components of the metabolic syndrome, including diabetes and obesity. Recently, several new obesity- and diabetes-related genes have been isolated and sequenced from *P. obesus*, including the beacon gene⁵⁵. Intracerebroventricular infusion of beacon protein results in a dose-dependent increase in food intake and body weight and an increase in hypothalamic expression of neuropeptide Y, a brain peptide that stimulates food intake.

In the other corner sit the supporters of the thrifty phenotype hypothesis, which is based on the epidemiological observations linking low birth weight with the risk of adult disease (obesity, diabetes and hypertension). Hales and Barker have hypothesized⁵⁶ that intrauterine malnutrition leads to reduced birth size and to permanent changes in structure and function, which predispose to disease in adult life. Although accepting the association between low birth weight and subsequent disease, a number of authors have questioned the 'environmental' interpretation, pointing out that, paradoxically, the data could be consistent with the thrifty genotype scenario⁷. Given the strong evidence for a role for genes in type 2 diabetes, it is also possible that the surviving low-birth-weight babies are actually an example of the thrifty gene operating — fetuses with low birth weight that survive are the example of a 'survival of the fittest' gene in an environment of intrauterine malnutrition⁵⁷. Because they have the thrifty genotype, these survivors would also be at risk of disease in later life.

More detailed reviews on this debate and a number of related issues are available elsewhere^{57,58}. The reviewers point out that claiming causal inference in this association must be seriously challenged. Programming studies are unique in that potential causes are temporally separated from effects by a span of some five decades. Joseph and Kramer⁵⁹ list various direct and indirect evidence that suggests the reported association of low birth weight and the later development of type 2 diabetes may be biased rather than causal. They state: "...selection bias, failure to define, measure, and adequately control for the confounding health consequences of social deprivation and inconsistencies in the hypotheses tested and in methods of data analysis and reporting are among the factors that weigh against a causal explanation for the associations observed". Even if the relationship is shown ultimately to be causal, low

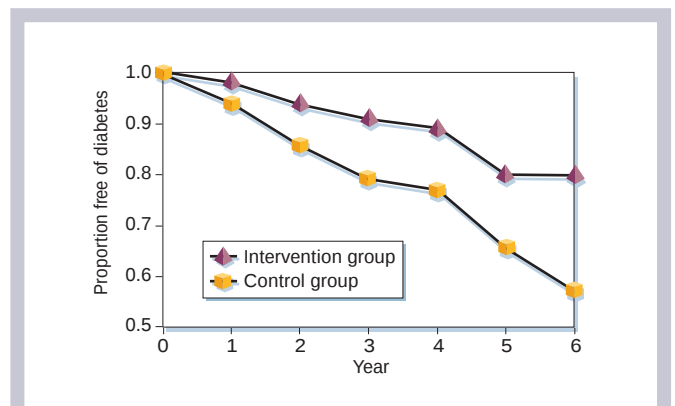


Figure 4 Reduction in risk of progressing from IGT to diabetes as a result of changes in intensive lifestyle. Data adapted with permission from ref. 22.

birth weight explains only a small proportion of diabetes, and does not exclude the importance of other hypotheses (such as the thrifty genotype).

Approaches to prevention of type 2 diabetes

Although focusing and directing funds at treating diabetes and its complications is important, the rapid escalation of numbers of people with diabetes demands urgent action on prevention^{60,61}. If not, the economic costs of premature morbidity and mortality from diabetes could absorb much of the health care budgets of both developing and affluent nations^{1,60}. Imagine the potential crisis in Asia, which will probably be home to 61% of the total global projected number of people with diabetes by 2010.

Recent studies have highlighted the potential for intervention in IGT subjects to reduce progression to type 2 diabetes. One such study is the recently completed Diabetes Prevention Program in the United States⁶². This showed that over three years, lifestyle intervention (targeting diet and exercise) reduced the risk of progressing from IGT to diabetes by 58%, whereas the oral hypoglycaemic drug, metformin, reduced risk by 31% (ref. 63). Two other large-scale studies — from Da Qing, China²¹, and from Finland²² — have also demonstrated the efficacy of lifestyle interventions. In the Finnish study, the cumulative incidence of diabetes after four years was 11% in the intervention group and 23% in the control group. During the trial, the risk of diabetes was reduced by 58% ($P < 0.001$) in the intervention group (Fig. 4), and was directly associated with changes in lifestyle. The same strategy is appropriate for a number of non-communicable diseases, but it is quite clear that for these strategies to work, there is a need for both a population approach and political commitment.

Several studies provide evidence that the approach aimed at high-risk individuals (for example, those with IGT) may not be enough to prevent all cases of type 2 diabetes. Data from the UKPDS indicate that pancreatic β -cell function is already substantially reduced at the time of clinical diagnosis of type 2 diabetes^{14,64}. Even at the earlier stage of IGT, β -cell function is already impaired⁶⁵ and intervention at this stage may be too late to prevent many cases of diabetes. In addition, intensive intervention programmes using well-developed behaviour-modification approaches often show high relapse rates⁶⁶, with weight gain and an increase in blood glucose occurring 1 to 2 years after an initially encouraging response. As a result, a number of studies are either underway or are being planned to examine whether therapeutic intervention with drugs such as metformin, α -glucosidase inhibitors, angiotensin-converting enzyme inhibitors and thiazolidinediones might be used for prevention as well.

What is not yet known is whether treatment of IGT and IFG specifically can delay or prevent the appearance of macrovascular disease, the main cause of morbidity and mortality in type 2 diabetes³⁰. However, delaying the onset of diabetes in such high-risk

subjects will itself provide benefit in terms of morbidity and mortality. It may therefore be prudent to treat such individuals with, at the least, lifestyle advice or with glucose-lowering agents of proven long-term safety while more data are accumulated.

Sociocultural perspectives of type 2 diabetes prevention

One of the myths of the modern world is that health is determined largely by individual choice⁶⁷. The myth is particularly exemplified in the area of NCDs such as type 2 diabetes. Changes in work patterns from heavy labour to sedentary, the increase in computerization and mechanization, and improved transport are just a few of the changes that have had an impact on human health¹. These should be considered in any approaches to improve health worldwide, along with decreasing easy access to fast foods and empty calories. Thus prevention requires a public health approach accompanied by major structural changes in society. Can we turn the clock back?

The strategy used by government officials in Philadelphia may not be the answer, but it certainly underlines the point. Philadelphia was voted fattest city in the United States and officials patrol the streets with weighing machines in an effort to persuade locals to lose weight⁶⁸. There is an urgent need to address the social and political issues in the public health approach to prevention of type 2 diabetes, some of which have been addressed recently by McKinlay and Marceau⁶⁹. Major shifts in public policy are necessary to create an environment for the whole community or nation in which individual behavioural initiatives can succeed. This may require changes in taxation and reimbursement for health promotion, provision of safe conditions (for example, from assault or traffic) for the elderly and younger sectors of the community, as well as community and workplace access to facilities for exercise.

Aetiology of type 1 diabetes

Type 1 diabetes is a discrete disorder and its pathogenesis involves environmental triggers that may activate autoimmune mechanisms in genetically susceptible individuals, leading to progressive loss of pancreatic islet β -cells⁷⁰. In many respects, attempts to solve the puzzle of the aetiology of type 1 diabetes have been disappointing despite decades of research^{70,71}. Predisposition is mediated by a number of genes that interact in a complex manner with each other and the environment^{70,71}.

Very low rates of type 1 diabetes have been found in Asian populations⁷², whereas Finland shows the world's highest incidence of 35 cases per 100,000 (ref. 73). This differs from other Baltic States⁷⁴, notably Estonia⁷⁵ whose population is linguistically and ethnically very similar to that of Finland, but who suffer only a third the incidence. This indicates that environmental factors have a particularly powerful influence on the appearance of type 1 diabetes.

Despite these compelling epidemiological findings, the environmental factor/s that precipitate type 1 diabetes in genetically susceptible individuals have remained speculative, although viruses — including *in utero* exposure, which is remote from clinical onset — have been suggested, as have dairy products and early weaning onto cow's milk, and increased dietary nitrate and nitrite (refs 70,76 and Table 3). The environmental agents that cause this injury are difficult to identify owing to the long period between exposure and the onset of hyperglycaemia, the complex genetics of the disease, and the likely multiple insults needed to initiate insulinitis.

We believe that the triggering agent is likely to be environmental and ubiquitous and have recently suggested a new possible candidate⁷⁷. We reported that minute quantities of the macrolide antibiotic bafilomycin A1 (bafA1) caused reduced glucose tolerance and disruption of the pancreatic islets in mice. BafA1 and related macrolides are produced by *Streptomyces* species. A *Streptomyces* species is also the source of streptozotocin, an agent used to produce experimental diabetes in rodents, either by a single administration or by multiple low-dose administration. The latter is taken to be a reliable model of induced autoimmune diabetes in rodents⁷⁸.

Table 3 Possible aetiological factors in type 1 diabetes

Non-genetic*	Genetic†
Viral infections (for example, coxsackie, cytomegalovirus)	Human leukocyte antigen (HLA) associated
Early infant diet (early cessation of breast feeding/early introduction of cow's milk)	Non-HLA associated
Perinatal infections	
Toxins (for example, dietary nitrosamines, bafilomycin, concanamycin A)	
Vaccine administration	

*No clear evidence for the role of any of these agents has been established.

†Consistent evidence for both HLA- and non-HLA-associated genes has been established.

Streptomyces species are ubiquitously present in soil and some can infest tuberous vegetables such as potatoes and sugar beet⁷⁷. Hence dietary exposure to a *Streptomyces* toxin could possibly cause repetitive pancreatic islet β -cell damage, and so be diabetogenic in humans genetically susceptible to autoimmune insulinitis.

Future perspectives

It will take a much more integrated and international approach to have a significant impact on the diabetes epidemic. We must accept that type 2 diabetes is not just a disease, but a symptom of a much larger global problem — the effect on human health of environmental and lifestyle changes¹.

It may not be too late to develop highly integrated policies for education and intervention^{69,76}. A large proportion of cases of type 2 diabetes is preventable. Initiatives for consumer education that promote a healthy diet could be reinforced by legislative changes such as increased taxation of certain 'unhealthy' foods. This is obviously a complex area, fraught with potential and political hazards.

Diabetes is likely to remain a huge threat to public health in the years to come. In the absence of effective and affordable (particularly for developing nations) interventions for both types of diabetes, the frequency will escalate worldwide, with the main impact being seen in developing nations and the disadvantaged minorities in developed nations^{2,3,45}. Thus prevention of diabetes and its micro- and macrovascular complications should be an essential component of future public health strategies for all nations. An urgent priority is the establishment of a multidisciplinary international task force representing all parties that can help reverse the underlying socioeconomic causes of the problem and address the issues that have led to the diabetes and NCD epidemic. □

1. Zimmet, P. Globalization, coca-colonization and the chronic disease epidemic: can the doomsday scenario be averted? *J. Intern. Med.* **247**, 301–310 (2000).
2. Amos, A., McCarty, D. & Zimmet, P. The rising global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetic Med.* **14**, S1–S85 (1997).
3. King, H., Aubert, R. & Herman, W. Global burden of diabetes, 1995–2025. Prevalence, numerical estimates and projections. *Diabetes Care* **21**, 1414–1431 (1998).
4. Astrup, A. & Finer, N. Redefining type 2 diabetes: 'diabesity' or 'obesity dependent diabetes mellitus'? *Obesity Rev.* **1**, 57–59 (2000).
5. Shafritz, E. Development and consequences of insulin resistance: lessons from animals with hyperinsulinaemia. *Diabetes Metab.* **22**, 131–148 (1997).
6. World Health Organization. Definition, Diagnosis and Classification of Diabetes mellitus and its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus (Department of Noncommunicable Disease Surveillance, Geneva, 1999).
7. Zimmet, P. Diabetes epidemiology as a trigger to diabetes research. *Diabetologia* **42**, 499–518 (1999).
8. Hennekens, G. & Buring, J. *Epidemiology in Medicine* (Little, Brown, Boston, 1987).
9. American Diabetes Association. Economic consequences of diabetes mellitus in the U.S. in 1997. *Diabetes Care* **21**, 296–309 (1998).
10. Fagot-Campagna, A. *et al.* Type 2 diabetes among North American children and adolescents: an epidemiologic review and a public health perspective. *J. Pediatr.* **136**, 664–672 (2000).
11. Fagot-Campagna, A. & Narayan, K. Type 2 diabetes in children. *Br. Med. J.* **322**, 377–387 (2001).
12. Songer, T. The economic costs of NIDDM. *Diabetes/Metabol. Rev.* **8**, 389–404 (1992).
13. Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.* **329**, 977–986 (1993).
14. United Kingdom Prospective Diabetes Study Group (UKPDS). Intensive blood-glucose control with sulfonylureas or insulin compared with conventional treatment and risk of complications in patients with Type 2 diabetes (UKPDS 33). *Lancet* **352**, 837–853 (1998).
15. Reichard, P., Nilsson, B.-Y. & Rosenqvist, U. The effect of long-term intensified insulin treatment on the development of microvascular complications of diabetes mellitus. *N. Engl. J. Med.* **329**, 304–309 (1993).

16. Heart Outcomes Prevention Evaluation (HOPE) Study. Effects of ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus: results of the HOPE study and MICRO-HOPE substudy. *Lancet* **355**, 253–259 (2000).
17. Ohkubo, Y. *et al.* Intensive insulin therapy prevents the progression of diabetic microvascular complications in Japanese patients with non-insulin-dependent diabetes mellitus: a randomized prospective 6-year study. *Diabetes Res. Clin. Pract.* **28**, 103–117 (1995).
18. Pyörälä, K., Pedersen, T., Kjekus, J., Faergeman, O. & Olsson, A. Cholesterol lowering with simvastatin improves prognosis of diabetic patients with coronary heart disease. A subgroup analysis of the Scandinavian Simvastatin Survival Study (4S). *Diabetes Care* **20**, 614–620 (1997).
19. Sacks, F., Pfeffer, M. & Moye, L. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and Recurrent Events Trial Investigators. *N. Engl. J. Med.* **335**, 1001–1009 (1996).
20. Zimmet, P. The pathogenesis and prevention of diabetes in adults. Genes, autoimmunity, and demography. *Diabetes Care* **18**, 1050–1064 (1995).
21. Pan, X. *et al.* Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. *Diabetes Care* **20**, 537–544 (1997).
22. Tuomilehto, J. *et al.* Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N. Engl. J. Med.* **344**, 1343–1350 (2001).
23. Eastman, R., Javitt, J., Herman, W., Dasbach, E. & Harris, M. in *Diabetes Mellitus: A Fundamental and Clinical Text* (eds Le-Roith, D., Taylor, S. & Olefsky, J.) 621–630 (Lippincott-Raven, New York, 1996).
24. World Health Organization. Diabetes Mellitus: Report of a WHO Study Group (WHO, Geneva, 1985).
25. Harris, M. & Zimmet, P. in *International Textbook of Diabetes Mellitus* 2nd edn (eds Alberti, K., Zimmet, P. & DeFronzo, R.) 9–23 (Wiley, Chichester, 1997).
26. Shaw, J., Hodge, A., de Courten, M., Chitson, P. & Zimmet, P. Isolated post-challenge hyperglycaemia confirmed as a risk factor for mortality. *Diabetologia* **42**, 1050–1054 (1999).
27. DECODE Study Group. Will new diagnostic criteria for diabetes mellitus change phenotype of patients with diabetes? Reanalysis of European epidemiological data. *Br. Med. J.* **317**, 371–375 (1998).
28. Tominaga, M. *et al.* Impaired glucose tolerance is a risk factor for cardiovascular disease, but not impaired fasting glucose. The Funagata Diabetes Study. *Diabetes Care* **22**, 920–924 (1999).
29. Perry, R. & Baron, A. Impaired glucose tolerance. Why is it not a disease? *Diabetes Care* **22**, 883–884 (1999).
30. Zimmet, P. & Alberti, K. The changing face of macrovascular disease in non-insulin dependent diabetes mellitus in different cultures: an epidemic in progress. *Lancet* **350**, S1–S4 (1997).
31. Alberti, K. G. M. M. Impaired glucose tolerance—fact or fiction. *Diabetic Med.* **13**, 6–8 (1996).
32. American Diabetes Association Expert Committee. Report of the Expert Committee. *Diabetes Care* **20**, 1183–1197 (1997).
33. Shaw, J. *et al.* Impaired fasting glucose or impaired glucose tolerance. What best predicts future diabetes in Mauritius. *Diabetes Care* **22**, 399–402 (1999).
34. Groop, L. The molecular genetics of non-insulin dependent diabetes mellitus. *J. Intern. Med.* **241**, 95–110 (1997).
35. Haffner, S. M., Stern, M. P., Hazuda, H. P., Mitchell, B. D. & Patterson, J. K. Cardiovascular risk factors in confirmed prediabetic individuals. Does the clock for coronary heart disease start ticking before the onset on clinical diabetes? *J. Am. Med. Assoc.* **263**, 2893–2898 (1990).
36. Reaven, G. Role of insulin resistance in human disease. *Diabetes* **37**, 1595–1607 (1988).
37. Zimmet, P., Boyko, E., Collier, G. & de Courten, M. in *The Metabolic Syndrome X* Vol. 892 (eds Hansen, B., Saye, J. & Wennogle, L.) 25–44 (New York Academy of Sciences, New York, 1999).
38. Isomaa, B. *et al.* Cardiovascular morbidity and mortality associated with the metabolic syndrome. *Diabetes Care* **24**, 683–689 (2001).
39. Rosenbloom, A., Joe, J., Young, R. & Winter, W. Emerging epidemic of Type 2 diabetes in youth. *Diabetes Care* **22**, 345–354 (1999).
40. American Diabetes Association. Type 2 diabetes in children and adolescents. *Diabetes Care* **23**, 381–389 (2000).
41. Ehtishami, S., Barrett, T. & Shaw, N. Type 2 diabetes mellitus in UK children—an emerging problem. *Diabetic Med.* **17**, 867–871 (2000).
42. Kitagawa, T., Owada, M., Urakami, T. & Yamachi, K. Increased incidence of non-insulin dependent diabetes mellitus among Japanese school children correlates with an increased intake of animal protein and fat. *Clin. Pediatr.* **37**, 111–116 (1998).
43. Dowse, G., Zimmet, P., Finch, C. & Collins, V. Decline in incidence of epidemic glucose intolerance in Nauruans: implications for the thrifty genotype. *Am. J. Epidemiol.* **133**, 1093–1104 (1991).
44. Zimmet, P., Turner, R., McCarty, D., Rowley, M. & Mackay, I. Crucial points at diagnosis—NIDDM or slow IDDM. *Diabetes Care* **22**, B59–B64 (1999).
45. de Courten, M., Bennett, P. H., Tuomilehto, J. & Zimmet, P. in *International Textbook of Diabetes Mellitus* 2nd edn (eds Alberti, K. G. M. M., Zimmet, P., DeFronzo, R. A. & Keen, H.) 143–170 (Wiley, Chichester, 1997).
46. Ramachandran, A., Snehalatha, C., Latha, E., Vijay, V. & Viswanathan, M. Rising prevalence of NIDDM in an urban population in India. *Diabetologia* **40**, 232–237 (1997).
47. O'Dea, K. Westernisation, insulin resistance and diabetes in Australian Aborigines. *Med. J. Aust.* **155**, 258–264 (1991).
48. Zimmet, P., Dowse, G., Finch, C., Serjeantson, S. & King, H. The epidemiology and natural history of NIDDM—lessons from the South Pacific. *Diabetes/Metabol. Rev.* **6**, 91–124 (1990).
49. Tan, C., Emmanuel, S., Tan, B.-Y. & Jacob, E. Prevalence of diabetes and ethnic differences in cardiovascular risk factors. *Diabetes Care* **22**, 241–247 (1999).
50. Chou, P., Chen, H. & Hsiao, K. Community-based epidemiological study on diabetes in Pu-Li, Taiwan. *Diabetes Care* **15**, 81–89 (1992).
51. Pan, X.-R., Yang, W.-Y., Li, G.-W. & Liu, J. Prevalence of diabetes and its risk factors in China. National Diabetes Prevention and Control Cooperative Group. *Diabetes Care* **20**, 1664–1669 (1997).
52. Zimmet, P. Kelly West Lecture 1991. Challenges in diabetes epidemiology—from West to the Rest. *Diabetes Care* **15**, 232–252 (1992).
53. Neel, J. Diabetes mellitus: a thrifty genotype rendered detrimental by “progress”? *Am. J. Hum. Genet.* **14**, 353–362 (1962).
54. Dowse, G. & Zimmet, P. The thrifty genotype in non-insulin-dependent diabetes. The hypothesis survives. *Br. Med. J.* **306**, 532–533 (1993).
55. Collier, G. *et al.* Beacon: A novel gene involved in the regulation of energy balance. *Diabetes* **49**, 1766–1771 (2000).
56. Hales, C. N. & Barker, D. J. P. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia* **35**, 595–601 (1992).
57. Harding, J. The nutritional basis of the fetal origins of adult disease. *Int. J. Epidemiol.* **30**, 15–23 (2001).
58. Hattersley, A. & Tooke, J. The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease. *Lancet* **353**, 1789–1792 (1999).
59. Joseph, K. & Kramer, M. Review of the evidence on fetal and early childhood antecedents of adult chronic disease. *Epidemiol. Rev.* **18**, 158–174 (1996).
60. Zimmet, P. & Lefebvre, P. The global NIDDM epidemic. Treating the disease and ignoring the symptom. *Diabetologia* **39**, 1247–1248 (1996).
61. Zimmet, P. & Chitson, P. Non-Communicable Diseases in *Health in the Commonwealth* 74–80 (Kensington Publications, London, 2000).
62. The Diabetes Prevention Program. Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care* **22**, 623–634 (1999).
63. Larkin, M. Diet and exercise delay onset of type 2 diabetes, say US experts. *Lancet* **358**, 565 (2001).
64. UKPDS. Overview of 6 years’ therapy of type II diabetes: a progressive disease (UKPDS 16). *Diabetes* **44**, 1249–1258 (1995).
65. Sicree, R., Zimmet, P., King, H. & Coventry, J. Plasma insulin response among Nauruans: prediction of deterioration in glucose tolerance over 6 yrs. *Diabetes* **36**, 179–186 (1987).
66. Mann, J. Stemming the tide of diabetes mellitus. *Lancet* **356**, 1454–1455 (2000).
67. Bloom, B. Public-health priorities in the industrialised world. *Lancet* **356**(Suppl. 1), 50–51 (2000).
68. Bonetta, L. Scientists battle obesity overload. *Nature Med.* **7**, 387 (2001).
69. McKinlay, J. & Marceau, L. US public health and the 21st century: diabetes mellitus. *Lancet* **356**, 757–761 (2000).
70. Harrison, L., Kay, T., Colman, P. & Honeyman, M. in *Diabetes in the New Millennium* (eds Turtle, J. & Osato, S.) 85–100 (The Endocrinology and Diabetes Research Foundation of the University of Sydney, Sydney, 1999).
71. Atkinson, M. & Maclaren, N. The pathogenesis of insulin-dependent diabetes mellitus. *N. Engl. J. Med.* **331**, 1428–1436 (1994).
72. Karvonen, M., Tuomilehto, J., Libman, I. & LaPorte, R. A review of the recent epidemiological data on the worldwide incidence of Type 1 (insulin-dependent) diabetes mellitus. *Diabetologia* **36**, 883–892 (1993).
73. Tuomilehto, J. *et al.* Epidemiology of childhood diabetes mellitus in Finland—background of a nationwide study of type 1 diabetes (insulin dependent) diabetes mellitus. *Diabetologia* **36**, 70–76 (1992).
74. Tuomilehto, J. *et al.* Comparison of the incidence of insulin dependent diabetes mellitus among five Baltic populations during 1983–1988. *Int. J. Epidemiol.* **21**, 518–527 (1992).
75. Tuomilehto, J. *et al.* Comparison of incidence of IDDM in childhood between Estonia and Finland, 1980–1988. *Diabetes Care* **14**, 982–988 (1991).
76. Tuomilehto, J., Tuomilehto-Wolf, E., Zimmet, P., Alberti, K. & Knowler, W. in *International Textbook of Diabetes Mellitus* 2nd edn (eds Alberti, K., Zimmet, P., DeFronzo, R. & Keen, H.) 1799–1827 (Wiley, Chichester, 1997).
77. Myers, M., Mackay, I., Rowley, M. & Zimmet, P. Dietary microbial toxins and Type 1 diabetes—a new meaning for seed and soil. *Diabetologia* **44**, 1199–1200 (2001).
78. Herold, K., Bloch, T., Vezys, V. & Sun, Q. Diabetes induced with low doses of streptozotocin is mediated by V beta 8.2+ cells. *Diabetes* **44**, 354–359 (1995).

Diabetes mellitus and genetically programmed defects in β -cell function

Graeme I. Bell & Kenneth S. Polonsky

Howard Hughes Medical Institute and Departments of Biochemistry and Molecular Biology, Medicine and Human Genetics, The University of Chicago, Chicago, Illinois 60637, USA (e-mail: g-bell@uchicago.edu)

Departments of Medicine, and Cell Biology and Physiology, Washington University School of Medicine, St. Louis, Missouri 63110, USA (e-mail: polonsky@im.wustl.edu)

The pathways that control insulin secretion and regulate pancreatic β -cell mass are crucial in the development of diabetes mellitus. Maturity-onset diabetes of the young comprises a number of single-gene disorders affecting pancreatic β -cell function, and the consequences of mutations in these genes are so serious that diabetes develops in childhood or adolescence. A genetic basis for the more common form of type 2 diabetes, which affects 10–20% of adults in many developed countries, is less clear cut. It is also characterized by abnormal β -cell function, but other tissues are involved as well. However, in both forms identification of causative and susceptibility genes are providing new insight into the control of insulin action and secretion, as well as suggesting new treatments for diabetes.

Diabetes mellitus is a heterogeneous group of disorders characterized by high blood glucose levels¹. The pancreatic β -cell and its secretory product insulin are central in the pathophysiology of diabetes. Type 1 or insulin-dependent diabetes mellitus results from an absolute deficiency of insulin due to autoimmune destruction of the insulin-producing pancreatic β -cell². In type 2 or non-insulin-dependent diabetes mellitus, muscle and fat cells are 'resistant' to the actions of insulin and compensatory mechanisms that are activated in the β -cell to secrete more insulin are not sufficient to maintain blood glucose levels within a normal physiological range^{3,4}.

Although the consequences of autoimmune destruction of pancreatic β -cells are clear, the causes are not, and the identification of the environmental factors that trigger this destruction in genetically susceptible children, adolescents and even adults still eludes us.

Type 2 diabetes is made up of different forms each of which is characterized by variable degrees of insulin resistance and β -cell dysfunction, and which together lead to hyperglycaemia¹. At each end of this spectrum are single-gene disorders that affect the ability of the pancreatic β -cell to secrete insulin^{5,6} or the ability of muscle, fat and liver cells to respond to insulin's actions^{7,8}.

Normal and altered insulin secretion

Insulin concentrations are normally determined by a feedback control system that is responsive to the prevailing level of plasma glucose⁹. The overall sensitivity of the pancreatic β -cell to glucose is determined by the sensitivity of peripheral tissues to the action of insulin with insulin-resistant subjects having higher insulin levels and insulin secretion rates than insulin-sensitive subjects. Insulin is also secreted in response to amino acids and fatty acids and the magnitude of this response being modulated by a variety of neural (for example, sympathetic and parasympathetic autonomic tone) and hormonal factors (for example, glucagon, glucagon-like peptide 1, gastric inhibitory polypeptide and somatostatin). Glucose, however, is the overriding influence.

Normal insulin secretion shows a rapid response to glucose and a complex pulsatile profile. The increase in insulin secretion that occurs after the intravenous administration of glucose is virtually instantaneous; even after oral glucose ingestion, increases in insulin secretion occur within minutes. The temporal profile of insulin secretion consists of small-amplitude pulses of insulin occurring every 5–10 minutes, superimposed on slower, larger-amplitude oscillations that occur every 1–2 hours.

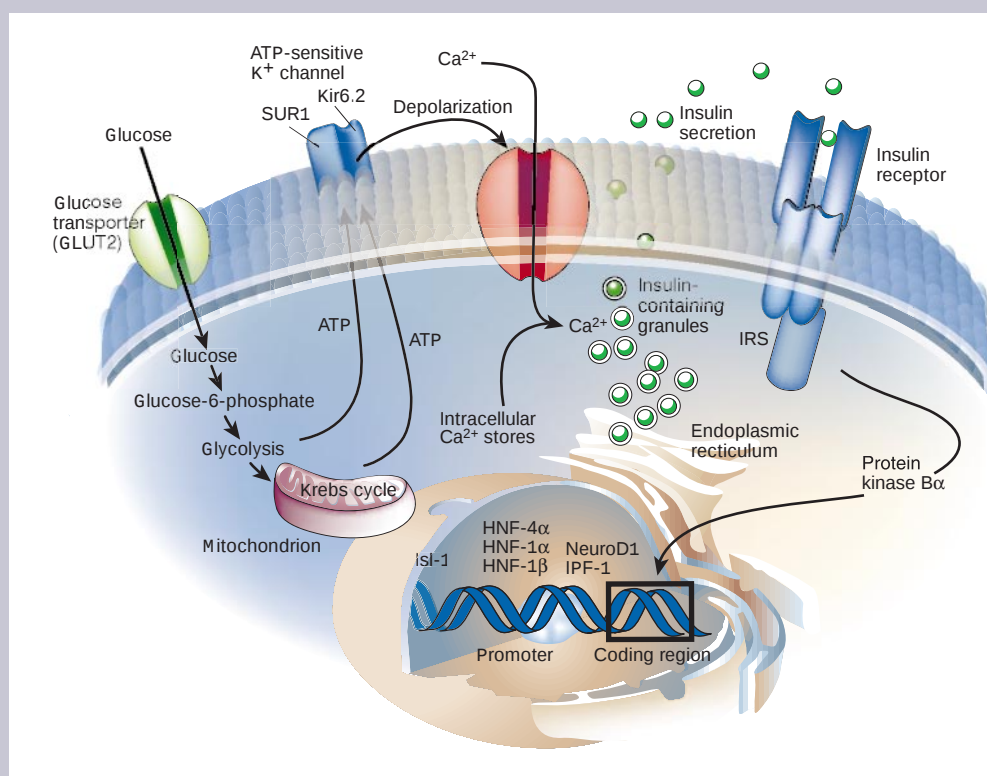
Early manifestations of disordered β -cell function include delayed and blunted responses to glucose, temporal irregularities in the pulses and oscillations of insulin secretion, and loss of the tight coupling between pulses of insulin secretion and pulses in glucose. Diabetes is also associated with impaired conversion of proinsulin to insulin. As a result, circulating proinsulin concentrations comprise a greater proportion of total immunoreactive insulin concentrations in diabetic than in normal individuals.

Genetic deficiencies in insulin secretion

Maturity-onset diabetes of the young (MODY) is a clinically heterogeneous group of disorders characterized by non-ketotic diabetes mellitus, an autosomal dominant mode of inheritance, onset usually before 25 years of age and frequently in childhood or adolescence, and a primary defect in pancreatic β -cell function^{5,6}. MODY can result from mutations in any one of at least six different genes (Fig. 1) that encode the glycolytic enzyme glucokinase and five transcription factors: hepatocyte nuclear factor (HNF)-4 α , HNF-1 α , insulin promoter factor-1 (IPF-1), HNF-1 β and neurogenic differentiation 1/ β -cell E-box transactivator 2 (NeuroD1/BETA2). All are expressed in the pancreatic β -cell and mutations lead to β -cell dysfunction and diabetes mellitus in the heterozygous state. They are also expressed in other tissues and abnormalities in liver and kidney function may occur. Non-genetic factors that affect insulin sensitivity such as infection, puberty, pregnancy and, rarely, obesity may trigger the onset of diabetes and affect the severity of hyperglycaemia in MODY.

Glucokinase is expressed at highest levels in the pancreatic β -cell and the liver¹⁰. It catalyses the transfer of

Figure 1 Model of a pancreatic β -cell showing proteins associated with MODY and other forms of β -cell dysfunction. Glucose is transported into the β -cell by the glucose transporter 2 isoform (GLUT2). By catalysing the transfer of phosphate from ATP to glucose to form glucose-6-phosphate, glucokinase (MODY2) functions as the glucose sensor of the β -cell. The generation of ATP by glycolysis and the Krebs cycle leads to closure of the ATP-sensitive K^+ channel — a hetero-octamer comprised of four subunits of the sulphonylurea 1 receptor (SUR1) and four subunits of the inwardly rectifying K^+ channel Kir6.2 (ref. 59). Mutations in these proteins are associated with familial persistent hyperinsulinaemia hypoglycaemia of infancy⁵⁹. The closing of the ATP-sensitive K^+ channel leads to depolarization of the plasma membrane and influx of extracellular calcium. Together with calcium mobilized from intracellular stores, this leads to fusion of insulin-containing secretory granules with the plasma membrane and the release of insulin into the circulation. The pancreatic β -cells have insulin receptors and there is evidence for an autocrine action of insulin on β -cell function, including transcription of the glucokinase and insulin genes. The MODY-associated transcription factors HNF-4 α (MODY1),



HNF-1 α (MODY3), HNF-1 β (MODY5), IPF-1 (MODY4) and NeuroD1 (MODY6) regulate the transcription of insulin and other β -cell genes. Mutations in islet-1 (Isl-1) may also lead to β -cell dysfunction. Protein kinase B α may be important in determining β -cell mass.

phosphate from ATP to glucose to generate glucose-6-phosphate: the first rate-limiting step in glucose metabolism. Glucokinase functions as the glucose sensor in the β -cell by controlling the rate of entry of glucose into the glycolytic pathway and its subsequent metabolism. In the liver, glucokinase affects the ability to store glucose as glycogen, particularly in the postprandial state. Heterozygous mutations leading to partial deficiency of glucokinase are associated with MODY^{5,6} and homozygous mutations resulting in complete deficiency of this enzyme lead to permanent neonatal diabetes mellitus¹¹. A combination of reduced glucose-induced insulin secretion from the pancreatic β -cell and reduced glycogen storage in the liver lead to an increase in plasma glucose concentrations.

The transcription factors HNF-1 α , HNF-1 β and HNF-4 α are involved in the tissue-specific regulation of gene expression in the liver, pancreatic β -cells and other tissues^{12,13}. HNF-1 α and HNF-1 β are members of the homeodomain-containing family of transcription factors and HNF-4 α is an orphan nuclear receptor. They comprise part of a network of transcription factors that controls gene expression during embryonic development and in the adult tissues in which they are co-expressed. In the pancreatic β -cell, they regulate the expression of insulin as well as proteins involved in glucose transport, glycolysis and mitochondrial metabolism, all of which are important in the regulation of insulin secretion^{13–18}. Mutations in these genes produce defects in insulin secretory responses to a variety of factors, particularly glucose, which are present before the onset of hyperglycaemia^{5,6}. Reduced glucagon secretory responses to arginine have also been observed in patients with HNF-4 α mutations, suggesting a broader role for this transcription factor in pancreatic islet function and/or development¹⁹. In mice lacking the gene for HNF-1 α , β -cell mass fails to increase despite the presence of hyperglycaemia, which suggests that this transcription factor may regulate β -cell mass²⁰.

IPF-1 is a homeodomain-containing transcription factor involved in pancreatic development^{21–23}, transcriptional regulation of a number of β -cell genes including insulin, glucokinase, islet amyloid polypeptide and glucose transporter 2 (GLUT2)²², and mediation of glucose-stimulated insulin gene transcription²⁴. Mutations in the heterozygous state are associated with MODY²⁵, whereas mouse and human homozygotes fail to develop a pancreas and suffer congenital diabetes mellitus²³. IPF-1 mutations have also been discovered in a small fraction of patients with typical adult-onset type 2 diabetes^{26,27}. Subjects with heterozygous mutations in IPF-1 demonstrated reduced insulin secretory responses to glucose and glucagon-like peptide-1, consistent with a defect in the signalling pathways that regulate secretion in the β -cell and/or a defect in β -cell mass²⁸.

The basic helix–loop–helix transcription factor NeuroD1/BETA2 is involved in the regulation of transcription of the insulin gene and is required for normal pancreatic islet development²⁹. Mutations in NeuroD1 are a rare cause of MODY and result in reduced serum insulin concentrations³⁰, either due to a signalling defect in the β -cell, a reduction in β -cell mass, or both.

Mutations in other β -cell transcription factors may also contribute to the development of MODY or a MODY-like disorder. A nonsense mutation has been found in the LIM-homeodomain transcription factor islet-1 in a Japanese family³¹. This mutation led to decreased activity and thus may be the cause of diabetes in this family.

In addition to mutations in the nuclear genome, abnormal mitochondrial function resulting from mutations in the mitochondrial genome can lead to diabetes³². The most common diabetes-associated mutation is an A-to-G transition in the mitochondrial tRNA^{Leu(UUR)} gene at base pair 3,243. This results in defects in insulin secretion including failure of glucose to prime the insulin secretory response and abnormal insulin secretory oscillations³³.

Mice with β -cell-specific disruption of mitochondria develop diabetes as a result of reduced β -cell stimulus–secretion coupling, followed by reduction in β -cell mass³⁴.

β -Cell dysfunction in type 2 diabetes

In subjects with MODY or mitochondrial diabetes, the elevation in blood glucose levels is due to genetically determined defects in processes that are critical for the normal signalling pathways in the β -cell. The effects of these mutations on insulin secretion are so severe that diabetes develops in people with normal or relatively normal insulin sensitivity and without the contribution of environmental factors. The mutations themselves are sufficient to cause diabetes with hyperglycaemia occurring well before 30 years of age.

In type 2 diabetes, more moderate abnormalities of secretion are present that cause glucose intolerance only if insulin resistance is also present. The genetic basis of β -cell dysfunction in this form of diabetes is more complex, involving both multiple interacting genes and environmental factors, which determine whether diabetes will develop and at what age. Despite a genetic predisposition, diabetes may never manifest, and hyperglycaemia, when it occurs, usually does so later in life (50+ years)³⁵, although there has been a disturbing increase in the prevalence of type 2 diabetes in children in recent years³⁶. Here no genetic abnormalities profoundly affect β -cell function, but rather the inability of the pancreatic β -cell to adapt to the reductions in insulin sensitivity that occur over a lifetime precipitates the onset of hyperglycaemia.

Obesity is the most common cause of insulin resistance in humans³. In obese subjects, insulin levels typically increase to maintain normal glucose tolerance. Basal and total 24-h rates of insulin secretion are three to four times higher in obese insulin-resistant subjects than in lean controls³. The hyperinsulinaemia associated with insulin resistance results from a combination of an increase in insulin secretion and a reduction in insulin clearance rates. As a result, more insulin is secreted at each level of glucose; how much more represents the extent of the β -cell adaptation necessary to maintain normal glucose tolerance. In addition, there is an inverse relationship between insulin sensitivity and insulin secretion, which is indicative of β -cell compensation for insulin resistance⁴. As insulin resistance increases, the early first-phase insulin response must increase proportionately to maintain normal glucose tolerance. But in type 2 diabetes, there is an overall reduction in insulin secretion rate as the β -cell can no longer secrete sufficient insulin to maintain normal blood glucose levels.

The compensatory hypersecretion of insulin in insulin-resistant states is due to an expansion of β -cell mass^{37,38} and alterations in the expression of key enzymes of β -cell glucose metabolism. A diabetes-prone sub-line of Zucker fatty rats fail to adequately increase their β -cell mass in the face of insulin resistance, highlighting this as an essential component of the compensatory mechanisms that maintain normal glucose tolerance. This is not due to a failure of β -cell proliferation, but rather an enhanced rate of β -cell death presumably due to apoptosis.

The hypersecretion of insulin observed in insulin-resistant states is believed to be a consequence of increased levels of the glycolytic enzyme hexokinase^{38,39}. In normal pancreatic β -cells, glucokinase mediates the conversion of glucose to glucose-6-phosphate and so determines the threshold at which glucose stimulates insulin secretion¹⁰. Insulin resistance is associated with increased β -cell hexokinase activity, an enzyme with more than 100-fold-greater affinity for glucose, leading to secretion of insulin at lower glucose concentrations.

These data imply that the hypersecretion of insulin in insulin-resistant states such as obesity is a consequence of both an increase in β -cell mass and hexokinase activity in these cells. Nevertheless, important questions related to the mechanisms responsible for this compensatory hypersecretion remain unanswered. A number of hormones and growth factors have been shown to stimulate β -cell

growth and proliferation including growth hormone, prolactin, glucagon-like peptide-1, insulin-like growth factors and parathyroid hormone-related protein⁴⁰. Insulin resistance states themselves may lead to the release of an unidentified circulating islet cell factor that can promote β -cell replication⁴¹. Overexpression of protein kinase B α /Akt1 in the pancreatic β -cell also leads to a marked increase in β -cell mass⁴². But which, if any, of these growth factors are responsible for the increase in β -cell growth with insulin resistance and by what mechanisms remains unknown.

In a high proportion of obese subjects, β -cell compensation maintains normal glucose tolerance for many decades, preventing the development of diabetes. The factors that determine the ability of the β -cell to compensate for insulin resistance are unknown; genetic factors probably are important, although as a sudden gain in weight can precipitate diabetes onset, nongenetic factors must also play a role.

It has been suggested that increased free fatty acids in serum could precipitate β -cell failure³. Short-term exposure of pancreatic islets to free fatty acids increases insulin secretion, but long-term exposure inhibits glucose-induced insulin secretion and biosynthesis, and may lead to β -cell death by apoptosis. These effects may be mediated by increased expression of proteins which uncouple glucose metabolism from oxidative phosphorylation — a key link between β -cell glucose metabolism and insulin secretion⁴³. Clinical studies have also indicated a role for free fatty acids in affecting β -cell function. The insulin resistance induced by an acute (90 min) elevation in free fatty acids is compensated for by an appropriate increase in insulin secretion, although the β -cell compensatory response was inadequate after chronic exposure (48 h)⁴⁴.

Genes and type 2 diabetes

The identification and characterization of the genes involved in type 2 diabetes will add an essential level to our understanding of the pathways regulating β -cell function, including those for β -cell compensation. A common amino-acid polymorphism (Pro12Ala) in peroxisome proliferator-activated receptor- γ (PPAR γ) has been associated with type 2 diabetes⁴⁵. People homozygous for the Pro12 allele are more insulin resistant than those having one Ala12 allele and have a 1.25-fold increased risk of developing diabetes. There is also evidence for interaction between this polymorphism and fatty acids, thereby linking this locus with diet^{46,47}.

PPAR γ is the target of a class of insulin-sensitizing drugs used to treat patients with type 2 diabetes — the thiazolidinediones⁴⁸. Treatment with PPAR γ agonists is associated with lower levels of insulin resistance and improved insulin secretory responses to glucose⁴⁹. The expression of PPAR γ in insulin-responsive tissues (fat and muscle)⁴⁸ and pancreatic β -cells⁵⁰ provides a link between insulin resistance and insulin secretion. Furthermore, the recent demonstration that insulin-mediated (that is, autocrine) signalling pathways are important in the preservation of normal β -cell function⁵¹ raises the possibility that insulin resistance in the β -cell, developing in parallel to insulin resistance in muscle, fat and liver, could contribute directly to β -cell dysfunction in type 2 diabetes. Diabetic islets show reduced insulin gene transcription; this might be due, at least in part, to reduced insulin action in that tissue, and indicates that activation of insulin gene transcription is an important effect of insulin-mediated signalling⁵².

Genetic variation in the gene encoding calpain-10, a ubiquitously expressed cysteine protease, has also been associated with type 2 diabetes, increasing risk as much as threefold⁵³ through effects on both the normal function of the β -cell and insulin action in muscle and fat. Nondiabetic subjects with the at-risk haplotype or genotype secrete significantly less insulin in response to a glucose challenge (M. Walker and G.I.B., unpublished data). They may be more insulin resistant with a decreased rate of insulin-mediated glucose turnover in muscle and fat⁵⁴. The results of *in vitro* studies using calpain inhibitors are consistent with a role for calpains in these processes⁵⁵.

The identification of the substrates for calpain-10 in β -cells, muscle and fat may lead to the identification of other susceptibility genes.

The identification of the genes that increase the risk of type 2 diabetes has been an arduous task and has proceeded slowly. However, with completion of genome-wide scans for susceptibility genes in many different populations⁵⁶ and the presence of a draft sequence of the human genome and a catalogue of human variation⁵⁷, we expect that more will be identified in the near future. Some will be candidate genes such as IPF-1 and PPAR γ , whereas others like calpain-10 will identify new biochemical pathways.

Looking ahead

Although the task of unravelling the mysteries of a disease as complex as type 2 diabetes is daunting, we are confident that success will be achieved. As additional genes involved in the pathogenesis of type 2 diabetes are identified, a clearer understanding of the mechanisms responsible for disease development will emerge. This will lead to new insights into the genetic epidemiology of type 2 diabetes and the nature of the gene–gene and gene–environment interactions. It will also facilitate more refined clinical studies in humans that will allow the specific phenotypic manifestations associated with different genotypes to be defined. These studies will be greatly aided by new approaches to evaluating subtle changes in phenotype such as the recent development of methods of measuring β -cell mass *in vivo*⁵⁸. Studies in appropriate animal models will provide additional insights into the physiological effects of specific susceptibility genes. Finally, we anticipate that a better understanding of the mechanisms involved in the compensation of β -cell function for insulin resistance and the factors responsible for the eventual decompensation will lead to the identification of new therapeutic targets for both the prevention and treatment of diabetes. It may even be possible to develop treatments that target general processes that are applicable to the treatment of many of the multiple forms of diabetes. □

1. American Diabetes Association. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* **24**(Suppl. 1), S5–S20 (2001).
2. Atkinson, M. A. & Eisenbarth, G. S. Type 1 diabetes: new perspective on disease pathogenesis and treatment. *Lancet* **358**, 221–229 (2001).
3. Cavaghan, M. K., Ehrmann, D. A. & Polonsky, K. S. Interactions between insulin resistance and insulin secretion in the development of glucose intolerance. *J. Clin. Invest.* **106**, 329–333 (2000).
4. Kahn, S. E. The importance of β -cell failure in the development and progression of type 2 diabetes. *J. Clin. Endocrinol. Metab.* **86**, 4047–4058 (2001).
5. Fajans, S. S., Bell, G. I. & Polonsky, K. S. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N. Engl. J. Med.* **345**, 971–980 (2001).
6. Owen, K. & Hattersley, A. T. Maturity-onset diabetes of the young: from clinical description to molecular genetic characterization. *Best Pract. Res. Clin. Endo. Metab.* **15**, 309–323 (2001).
7. Taylor, S. I. & Arioglu, E. Genetically defined forms of diabetes in children. *J. Clin. Endocrinol. Metab.* **84**, 4390–4396 (1999).
8. Barroso, I. *et al.* Dominant negative mutations in human PPAR γ associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* **402**, 880–883 (1999).
9. Polonsky, K. S. & O'Meara, N. M. In *Endocrinology* 4th edn (eds DeGroot, L. J. & Jameson, L. J.) 697–711 (Saunders, Philadelphia, 2001).
10. Newgard, C. G. & Matschinsky, F. M. In *The Endocrine Pancreas and Regulation of Metabolism* (eds Jefferson, L. S. & Cherrington, A. D.) 125–151 (Oxford Univ. Press, Oxford, 2001).
11. Njolstad, P. R. *et al.* Neonatal diabetes mellitus due to complete glucokinase deficiency. *N. Engl. J. Med.* **344**, 1588–1592 (2001).
12. Cereghini, S. Liver-enriched transcription factors and hepatocyte differentiation. *FASEB J.* **10**, 267–282 (1996).
13. Ryffel, G. U. Mutations in the human genes encoding the transcription factors of the hepatocyte nuclear factor (HNF) 1 and HNF 4 families: functional and pathological consequences. *J. Mol. Endocrinol.* **27**, 11–29 (2001).
14. Emens, L. A., Landers, D. W. & Moss, L. G. Hepatocyte nuclear factor 1 α is expressed in a hamster insulinoma line and transactivates the rat insulin 1 gene. *Proc. Natl Acad. Sci. USA* **89**, 7300–7304 (1992).
15. Okita, K. *et al.* Human insulin gene is a target gene of hepatocyte nuclear factor-1 α (HNF-1 α) and HNF-1 β . *Biochem. Biophys. Res. Commun.* **263**, 566–569 (1999).
16. Wang, H., Antozzi, P. A., Hagenfeldt, K. A., Maechler, P. & Wollheim, C. B. Molecular targets of a HNF1 α mutation responsible for pancreatic β -cell dysfunction. *EMBO J.* **19**, 4257–4264 (2000).
17. Wang, H., Maechler, P., Antozzi, P. A., Hagenfeldt, K. A. & Wollheim, C. B. Hepatocyte nuclear factor 4 α regulates the expression of pancreatic β -cell genes implicated in glucose metabolism and nutrient-induced insulin secretion. *J. Biol. Chem.* **275**, 35953–35959 (2000).
18. Shih, D. Q. *et al.* Loss of HNF-1 α function in mice leads to abnormal expression of genes involved in pancreatic islet development and metabolism. *Diabetes* **50**, 2472–2480 (2001).
19. Herman, W. H. *et al.* Diminished insulin and glucagon secretory responses to arginine in nondiabetic subjects with a mutation in the hepatocyte nuclear factor-4 α /MODY1 gene. *Diabetes* **46**, 1749–1754 (1997).
20. Pontoglio, M. *et al.* Defective insulin secretion in hepatocyte nuclear factor 1 α -deficient mice. *J. Clin. Invest.* **101**, 2215–2222 (1998).

21. Jonsson, J., Carlsson, L., Edlund, T. & Edlund, H. Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* **371**, 606–609 (1994).
22. Edlund, H. Transcribing diabetes. *Diabetes* **47**, 1817–1823 (1998).
23. Stoffers, D. A., Zinkin, N. T., Stanojevic, V., Clarke, W. L. & Habener, J. F. Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nature Genet.* **15**, 106–110 (1997).
24. Marshak, S., Totary, H., Cerasi, E. & Melloul, D. Purification of the β -cell glucose-sensitive factor that transactivates the insulin gene differentially in normal and transformed islet cells. *Proc. Natl Acad. Sci. USA* **93**, 15057–15062 (1996).
25. Stoffers, D. A., Ferrer, J., Clarke, W. L. & Habener, J. F. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nature Genet.* **17**, 138–139 (1997).
26. Macfarlane, W. M. *et al.* Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *J. Clin. Invest.* **104**, R33–R39 (1999).
27. Hani, E. H. *et al.* Defective mutations in the insulin promoter factor-1 (IPF-1) gene in late-onset type 2 diabetes mellitus. *J. Clin. Invest.* **104**, R41–R48 (1999).
28. Clocquet, A. R. *et al.* Impaired insulin secretion and increased insulin sensitivity in familial maturity-onset diabetes of the young 4 (insulin promoter factor 1 gene). *Diabetes* **49**, 1856–1864 (2000).
29. Chu, K., Nemoz-Gaillard, E. & Tsai, M. J. BETA2 and pancreatic islet development. *Recent Prog. Horm. Res.* **56**, 23–46 (2001).
30. Malecki, M. T. *et al.* Mutations in NEUROD1 are associated with the development of type 2 diabetes mellitus. *Nature Genet.* **23**, 323–328 (1999).
31. Shimomura, H. *et al.* Nonsense mutation of islet-1 gene (Q310X) found in a type 2 diabetic patient with a strong family history. *Diabetes* **49**, 1597–1600 (2000).
32. Maassen, J. A. & Kadowaki, T. Maternally inherited diabetes and deafness: a new diabetes subtype. *Diabetologia* **39**, 375–382 (1996).
33. Velho, G. *et al.* Clinical phenotypes, insulin secretion, and insulin sensitivity in kindreds with maternally inherited diabetes and deafness due to mitochondrial tRNA^{Leu(UUR)} gene mutation. *Diabetes* **45**, 478–487 (1996).
34. Silva, J. P. *et al.* Impaired insulin secretion and β -cell loss in tissue-specific knockout mice with mitochondrial diabetes. *Nature Genet.* **26**, 336–340 (2000).
35. National Institutes of Health. *Diabetes in America* 2nd edn. NIH Publication No. 95–1468 (1995).
36. Ludwig, D. S. & Ebbeling, C. B. Type 2 diabetes mellitus in children. *J. Am. Med. Assoc.* **268**, 1427–1430 (2001).
37. Pick, A. *et al.* Role of apoptosis in failure of β -cell mass compensation for insulin resistance and β -cell defects in the male Zucker diabetic fatty rat. *Diabetes* **47**, 358–364 (1998).
38. Milburn, J. L. Jr *et al.* Pancreatic β -cells in obesity. Evidence for induction of functional, morphological, and metabolic abnormalities by increased long chain fatty acids. *J. Biol. Chem.* **270**, 1295–1299 (1995).
39. Cockburn, B. N. *et al.* Changes in pancreatic islet glucokinase and hexokinase activities with increasing age, obesity and the onset of diabetes. *Diabetes* **46**, 1434–1439 (1997).
40. García-Ocaña, A. *et al.* Using β -cell growth factors to enhance human pancreatic islet transplantation. *J. Clin. Endocrinol. Metab.* **86**, 984–988 (2001).
41. Flier, S. N., Kulkarni, R. N. & Kahn, C. R. Evidence for a circulating islet growth factor in insulin-resistant states. *Proc. Natl Acad. Sci. USA* **98**, 7475–7480 (2001).
42. Tuttle, R. L. *et al.* Regulation of pancreatic β -cell growth and survival by the serine/threonine protein kinase Akt/PKB α . *Nature Med.* **7**, 1133–1137 (2001).
43. Zhang, C. Y. *et al.* Uncoupling protein-2 negatively regulates insulin secretion and is a major link between obesity, β cell dysfunction, and type 2 diabetes. *Cell* **105**, 745–755 (2001).
44. Carpentier, A., Mittelman, S. D., Bergman, R. N., Giacca, A. & Lewis, G. F. Prolonged elevation of plasma free fatty acids impairs pancreatic β -cell function in obese nondiabetic human but not in individuals with type 2 diabetes. *Diabetes* **49**, 399–408 (2000).
45. Altshuler, D. *et al.* The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
46. Luan, J. *et al.* Evidence for gene–nutrient interaction at the PPAR γ locus. *Diabetes* **50**, 686–689 (2001).
47. Stefan, N., Fritsche, A., Häring, H. & Stumvoll, M. Effect of experimental elevation of free fatty acids on insulin secretion and insulin sensitivity in healthy carriers of the Pro12Ala polymorphism of the peroxisome proliferator-activated receptor- γ gene. *Diabetes* **50**, 1143–1148 (2001).
48. Olefsky, J. M. Treatment of insulin resistance with peroxisome proliferator-activated receptor γ agonists. *J. Clin. Invest.* **106**, 467–472 (2000).
49. Cavaghan, M. K., Ehrmann, D. A., Byrne, M. M. & Polonsky, K. S. Treatment with the oral antidiabetic agent troglitazone improves β cell responses to glucose in subjects with impaired glucose tolerance. *J. Clin. Invest.* **100**, 530–537 (1997).
50. Dubois, M. *et al.* Expression of the peroxisome proliferator-activated receptor γ (PPAR γ) in normal human pancreatic islets. *Diabetologia* **43**, 1165–1169 (2000).
51. Kulkarni, R. N. *et al.* Tissue-specific knockout of the insulin receptor in pancreatic β cells creates an insulin secretory defect similar to that in type 2 diabetes. *Cell* **96**, 329–339 (1999).
52. Leibiger, B. *et al.* Selective insulin signaling through A and B insulin receptors regulates transcription of the insulin and glucokinase genes in pancreatic β cells. *Molecular Cell* **7**, 559–570 (2001).
53. Horikawa, Y. *et al.* Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175 (2000).
54. Baier, L. J. *et al.* A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. *J. Clin. Invest.* **106**, R69–R73 (2000).
55. Sreenan, S. K. *et al.* Calpains play a role in insulin secretion and action. *Diabetes* **50**, 2013–2020 (2001).
56. Wilshire, S. *et al.* A genome-wide scan for loci predisposing to type 2 diabetes in the UK population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am. J. Hum. Genet.* **69**, 553–569 (2001).
57. National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov/> (2001).
58. Moore, A., Bonner-Weir, S. & Weissleder, R. Noninvasive *in vivo* measurement of β -cell mass in mouse model of diabetes. *Diabetes* **50**, 2231–2236 (2001).
59. Seino, S., Iwanaga, T., Nagashima, K. & Miki, T. Diverse roles of K_{ATP} channels learned from Kir6.2 genetically engineered mice. *Diabetes* **49**, 311–318 (2000).

Acknowledgements

We thank the members of our laboratories for their contributions to the studies presented here. This work was supported by grants from the NIH and a gift from the Blum-Kovler Foundation. G.I.B. is an investigator of the Howard Hughes Medical Institute.

β -Cell death during progression to diabetes

Diane Mathis, Luis Vence & Christophe Benoist

Section on Immunology and Immunogenetics, Joslin Diabetes Centre; Department of Medicine, Brigham and Women's Hospital; Harvard Medical School, One Joslin Place, Boston, Massachusetts 02215, USA (e-mail: cbdm@joslin.harvard.edu)

The hallmark of type 1 diabetes is specific destruction of pancreatic islet β -cells. Apoptosis of β -cells may be crucial at several points during disease progression, initiating leukocyte invasion of the islets and terminating the production of insulin in islet cells. β -Cell apoptosis may also be involved in the occasional evolution of type 2 into type 1 diabetes.

Type 1 diabetes is an autoimmune disease resulting from specific destruction of the insulin-producing β -cells of the islets of Langerhans of the pancreas¹. It has two distinct phases: insulinitis, when a mixed population of leukocytes invades the islets; and diabetes, when most β -cells have been killed off, and there is no longer sufficient insulin production to regulate blood glucose levels, resulting in hyperglycaemia. Individuals can have covert insulinitis for a long time (years in humans, months in rodent models) before it finally progresses to overt diabetes, and sometimes it never does.

Type 1 diabetes is an old disorder — descriptions of it appear in ancient Egyptian and Greek writings. It is also a common disease, currently affecting about 0.5% of the population in developed countries and increasing in incidence. In addition, there is mounting evidence that a fraction (variously estimated at 5–15%) of people originally diagnosed as type 2 diabetic may actually have a slowly progressing and less severe form of type 1 termed 'latent autoimmune diabetes of adults'. It is surprising, then, that we remain so ignorant about the aetiology and pathogene-

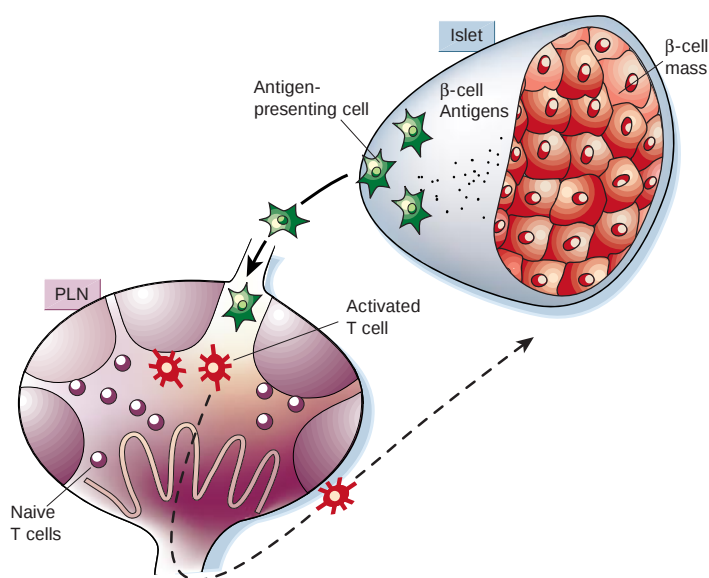
sis of autoimmune diabetes. We do not know what triggers it, have little understanding of the genetic and environmental factors regulating its progression, and have a confused view of the final effector mechanism(s). Consequently, we still do not know how to prevent or reverse type 1 diabetes in a sufficiently innocuous manner to be therapeutically useful. Faced with the difficulties of addressing these issues in humans, many investigators have turned to small-animal models of diabetes, in particular the non-obese diabetic (NOD) mouse and transgenic derivatives of it (Box 1).

Here we focus on the crux of autoimmune diabetes — β -cell death. We explore the possibility that physiological destruction of β -cells is a crucial event at disease outset, initiating autoimmunity. We also evaluate what is known about the mechanism(s) of β -cell death in the terminal phases of the autoimmune attack, and weigh the contention that destruction of β -cells can convert type 2 into type 1 diabetes.

Death as the initiator

Type 1 diabetes is primarily a T-lymphocyte-mediated disease^{1,2}. Immunohistological analyses show that most

Figure 1 Proposed scheme for the initiation of type 1 diabetes. Naive T cells (purple spheres) circulate through the blood and lymphoid organs, including the pancreatic lymph nodes (PLNs). In the nodes, they encounter APCs (probably mature DCs) displaying on their surface MHC molecules carrying antigens in the form of peptide fragments. In this case, the antigens derive from proteins synthesized by pancreatic islet β -cells, picked up (in soluble form, as cell bits or as apoptotic cells) when the APCs (probably DCs in immature state) resided in the islets. A minute fraction of the naive T cells can recognize MHC molecule/ β -cell antigen complexes, become activated (red asterisks), and then access tissues including the pancreas, where they re-encounter cognate antigen, and are reactivated and retained.



leukocytes in the islet infiltrate are T cells. Disease does not develop in NOD mice that are genetically athymic or T lymphopenic, or were thymectomized at birth; likewise, it is dampened or even abrogated by reagents that interfere with T-cell function. Finally, diabetes can be transferred by injecting T cells from diseased donors into healthy NOD recipients — the purest demonstration being inoculation of a single T-cell clone into a lymphocyte-deficient NOD mouse. Although most of these data derive from rodent diabetes models, there is ample indication that the human disease is also mediated by T lymphocytes, and has a similar islet histology and positive response to treatment with T-cell inhibitors.

T cells with diabetogenic properties fall into both the CD4⁺ helper and the CD8⁺ killer classes, although both may be required for optimum disease activity^{1,2}. They respond to several antigens that are specifically, but perhaps not uniquely, synthesized by pancreatic islet β -cells, including peptides derived from insulin, glutamic acid decarboxylase and the tyrosine phosphatases 1A-2 and 1A-2 β (phogrin). For CD4⁺ T cells, the antigens are peptides presented by major histocompatibility complex (MHC) class II molecules found primarily on specialized antigen-presenting cells (APCs); for CD8⁺ T cells, they are peptides displayed by MHC class I molecules present on most cell types.

Where and when the autoimmune response is initiated

Where do naive, potentially diabetogenic T cells first encounter their β -cell-derived antigens? The most likely possibilities are in the islets themselves; in the lymph nodes draining the islets (the pancreatic lymph nodes; PLNs); or at some distant site (although in this case the original stimulus would not be the β -cell antigen itself, but rather a molecular mimic of it introduced by infection or diet). Data from two transgenic mouse models of diabetes^{3,4} support the second explanation, suggesting a scheme (Fig. 1) that is highly analogous to that established for T-cell responses to antigens introduced through the skin.

Like all naive T lymphocytes, naive β -cell-reactive, potentially diabetogenic T cells do not usually access the tissues (including the pancreas), but circulate freely through the blood and lymphoid organs. At some point, β -cell-derived antigens are taken up by APCs (probably dendritic cells; DCs) in the islets, inducing maturation of the APCs and their migration to the immediately draining PLNs. Here, they display the β -cell-derived antigens to naive β -cell-reactive T cells in circulation and activate them. As part of the activation process, the T cells acquire the ability to migrate through the tissues; in the islets, they re-encounter cognate antigen, become reactivated and are retained, thereby initiating insulinitis. An unknown fraction of the activated and/or re-activated T cells may re-enter the circulation. Much evidence suggests that a similar sequence of events unfolds to initiate insulinitis in standard NOD mice³.

This scheme raises the issue of when β -cell antigens are first made available for uptake by islet-resident APCs. An answer has come from a T-cell receptor (TCR) transgenic mouse model of diabetes³. Insulinitis begins abruptly and highly synchronously in these animals at 15–18 days of age. At and after this time, it is possible to detect the relevant β -cell-derived antigen in the PLNs (and in no other lymph nodes) by monitoring activation and proliferation of transferred transgenic T cells; before this time, the PLNs seem to be free of the relevant antigen even though it is synthesized by β -cells from birth⁵. There seems to be no difference in the timing or extent of β -cell antigen release to the PLNs in NOD versus non-diabetes-prone C57Bl/6 mice⁶, implying that this event, albeit crucial to the process, does not explain the propensity of NOD mice to develop diabetes.

How autoimmunity is initiated

In regard to how β -cell antigens are first made available, an intriguing possibility is that they are taken up by APCs in the islets because of the death of β -cells. Many groups have shown that a wave of physiological β -cell death takes place in the islets of juvenile rodents, peaking at

Box 1

The NOD mouse and simpler murine models

Discovered in the late 1970s, the non-obese diabetic (NOD) mouse has become the most commonly used animal model of autoimmune diabetes². Disease develops spontaneously in this strain, sharing several crucial features with the human disorder — for example, a protracted course of pathology, disease that is primarily mediated by T lymphocytes, and disease that is under complex polygenic control, by far the most important contributor being the MHC. The NOD mouse has permitted many of the outstanding issues in the diabetes field to be addressed by direct experimentation, and results from this animal have heavily coloured our view of the pathogenesis of diabetes. But, because of the asynchrony and complexity of the disease in this model, vital issues have remained unresolved. Thus, diverse strategies have been taken to simplify the NOD disease in order to highlight particular features by ablation or exaggeration.

One strategy has been the generation of T-cell receptor (TCR) transgenic mice to focus on the role of T lymphocytes. Because of allelic exclusion, mice carrying already rearranged TCR transgenes have a T-cell repertoire highly skewed for the transgene-encoded specificity — even monoclonal when mutations blocking rearrangement of endogenous TCR genes (SCID, RAG^{0/0}) are introduced. Such a skewed repertoire has proved useful for synchronizing disease unfolding and for monitoring the behaviour of pathogenic T cells. There are now several examples of TCR transgenic mice with repertoires enriched for a diabetogenic specificity^{98,99} — either MHC class-I-restricted CD8⁺ or class-II-restricted CD4⁺ T cells, and T cells recognizing either a natural antigen or a genetically engineered neo-antigen synthesized specifically by β -cells.

Another strategy has been to direct expression of molecules of interest specifically to islet β -cells under the dictates of the insulin promoter, permitting assessment of the effects of expression at enhanced levels, ectopically, locally or in isolation. Examples include mice expressing MHC molecules, cytokines, co-stimulatory molecules and neo-antigens^{99,100}.

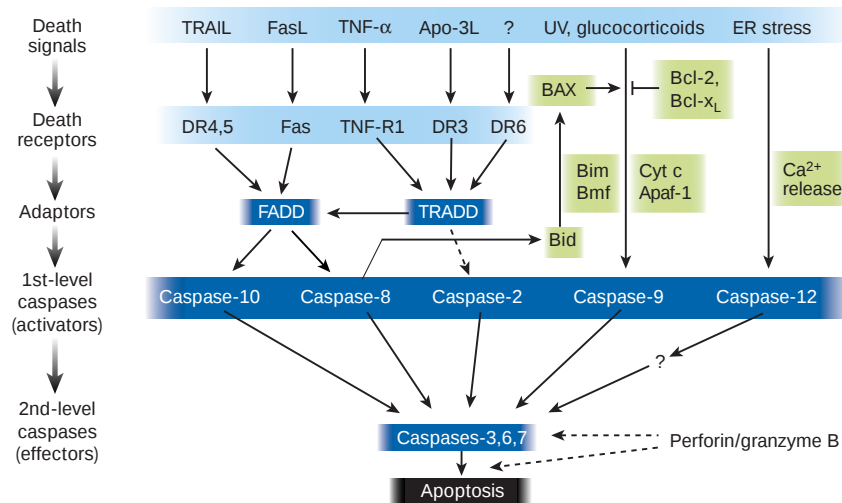
Thus, there now exists a set of powerful murine models of diabetes, each one with its inherent interests and disadvantages. We do not know, however, which model most faithfully mimics which variant of human diabetes. Given the heterogeneous symptoms, courses and genetics exhibited by affected individuals, it may well be that human diabetes is really a multiplicity of diseases. It seems that an open-minded view is called for at this point.

14–17 days after birth^{7–10}. Significant death at this age was first postulated on the basis of mathematical modelling studies, taking into account the kinetics of β -cell division, changes in cell volume, and the evolution of cell mass⁷. Confirmation came from quantifying dying cells on pancreas sections^{8,9}: the frequency of dying cells was low even at peak times (4–13%), but, given that dead cells can be very difficult to detect owing to rapid engulfment and disposal, it is possible that a significant fraction (estimated at one-third) of the β -cells in juvenile rodents turns over. A wave of β -cell death was observed in both normal and diabetes-prone strains, although perhaps augmented in some of the latter⁷. Interestingly, an analogous perinatal wave, peaking at birth, has been demonstrated in humans, leading to the suggestion that extension of this wave might be the cause of hyperinsulinaemia in infancy¹¹.

The significance of such a wave of β -cell death in the context of initiation of autoimmunity is that the engulfment of apoptotic and necrotic cells can be a potent activation signal for APCs, in particular DCs¹². Most recently, emphasis has been placed on necrotic cells as an immunogenic (as opposed to a tolerogenic) stimulus, but apoptotic

Figure 2 Intracellular death signalling molecules. Apoptosis is the end result of a cascade of interactions and activities⁹⁷. Mammals have two main networks of apoptosis signalling molecules, which are largely distinct. One is initiated by death receptors at the cell surface and is not controlled by Bcl-2 family members; the other is initiated by pro-apoptotic BH3-only members of the Bcl-2 family (such as Bid, Bim or Bmf), and is inhibited by anti-apoptotic Bcl-2 homologues (such as Bcl-2, Bcl-x_L). When cell-surface receptors such as Fas, TNF-R1 and death receptors (DR) 3–6 bind their ligands, the receptors cluster and signalling complexes are assembled, including adaptors (FADD, TRADD) and first-level pro-caspases (such as caspase-8 and -2);

and caspase-10 in humans). These pro-caspases are cleaved to generate active caspase forms, which activate downstream caspases, which in turn cleave several cell substrates to effect cell death. Other pathways initiated by ultraviolet light, glucocorticoids and stress, for example, do not involve cell-surface receptors, but rather mitochondrial perturbations and/or Ca²⁺ release. Other first-level caspases (caspase-9 and -12) are mobilized, followed by poorly characterized downstream events. Central to mitochondrial involvement are pro- and anti-apoptotic members of



the Bcl family, which regulate exit of cytochrome c, culminating in the assembly of an 'apoptosome'. The perforin/granzyme B pathway is also poorly defined, although at some point caspase-3 is mobilized. Perforin/granzyme B may also directly cleave cellular substrates. Dashed arrows indicate uncertainty, a controversial link, or one with limited experimental underpinning. No arrows emanate from Bmf and Bim because the pathways are currently ill defined.

cells can also provoke an immune response under the appropriate conditions, for example, when present in high enough numbers, physiologically stressed, or exposed to inflammatory cytokines such as tumour-necrosis factor (TNF)- α and interleukin (IL)- β (ref. 13). In addition, several studies have linked the presentation of autoantigens and production of autoantibodies with apoptotic cells, notably in systemic lupus erythematosus¹⁴. Thus, it seems plausible that this wave of β -cell death prompts uptake of β -cell antigens in the islets followed by their ferrying to the PLNs, where they become available to naive circulating β -cell-reactive T cells. In support of this, an earlier wave of apoptosis in the kidney (peak at 10 days) is correlated with an earlier arrival of kidney-derived antigens in the renal lymph nodes (at 10 days)^{3,10,15}.

The wave of β -cell death in juvenile animals might stem from one or both of two processes. First, at various times during development, certain tissues undergo a remodelling process involving apoptosis¹⁶ and/or necrosis¹⁷. A classic example is the loss of interdigital tissue in mammalian limbs during fetal development; examples in the neonatal window include death in the kidney, nervous system, heart and adrenal cortex. The precise mechanisms involved are poorly understood, but both cell-autonomous and extracellular controls have been implicated in different contexts, and the latter can involve either the imposition of a positive signal or the removal of a negative one¹⁸. Second, the wave of β -cell death peaks around the time of weaning — a known stage of massive metabolic and synthetic perturbations such as changes in glycogenolysis, gluconeogenesis, lipogenesis and hormone production, including insulin and glucagon¹⁹. Such changes might be an extracellular signal for triggering tissue remodelling, and/or they might induce tissue stress. The latter is likely to be relevant to the pancreatic islets, as β -cells are unusually sensitive to overproduction of proteins and to exposure to several cytokines²⁰.

The signalling networks involved in cell-stress-induced apoptosis are beginning to be identified. One major pathway, distinct from that implicated in death mediated by members of the TNF receptor family, mobilizes cytochrome c, Apaf1, caspase-9, and in some cases caspase-3 (refs 21–25; and Fig. 2). Yuan and co-workers²⁶ reported

that stress in the endoplasmic reticulum (ER), usually provoked by an excess of misfolded proteins, specifically activates caspase-12, which eventually leads to apoptosis. Cells from caspase-12-deficient mice are resistant to ER-stress-induced death, but do undergo apoptosis in response to other death stimuli²⁶. Also involved in this pathway are IRE-1, TRAF-2, calpain, caspase-7 and perhaps caspase-9 (refs 27–29; and Fig. 2). It is imperative to assess these two pathways in the context of insulinitis onset.

Thus, we and others⁷ propose that a wave of physiological β -cell death, which takes place in all individuals, is crucial in the initiation of type 1 diabetes. It has pathological consequences in diabetes-prone individuals, who have an autoreactive T-cell repertoire that can be mobilized by the exposed β -cell antigens and may also be subject to additional genetic or environmental susceptibility factors. This resulting indiscriminate release of β -cell antigens might explain why it has been so difficult to identify convincingly 'the antigen'³⁰ that initiates autoimmune diabetes. It also eliminates the need for invoking molecular mimicry by microbial T-cell epitopes to initiate the autoimmune process — a notion that has attracted much interest but has little direct experimental support³¹ — although microbes might still precipitate insulinitis prematurely or cause it to convert more rapidly to diabetes.

Proof will require juvenile β -cell death to be linked experimentally to onset of insulinitis; that is, by blocking the death wave or inducing it prematurely, we should detect a reduced or earlier leukocytic infiltrate, respectively. It may also prove possible to track apoptotic β -cell fragments as they are engulfed by APCs in the pancreas and transported to the PLNs, as has been done so elegantly in an intestinal system³². Finally, it will be important to define the death signalling pathways that are mobilized during the wave of β -cell death in juvenile animals.

Death as a terminator

Massive, specific destruction of pancreatic islet β -cells is the hallmark of type 1 diabetes. Thus, it is surprising that we still have so little definitive information — either from rodent models or from affected

humans — on how this comes about or why it is so specific. For humans, this most probably reflects the fact that once diabetes is diagnosed, most of the β -cell destruction has already taken place.

Cellular mechanisms of β -cell death

Two principal cellular mechanisms have been proposed (refs 1,2; and Fig. 3). The 'recognition-linked' mechanism attributes the death of β -cells to cytotoxic T-cell recognition of autoantigens presented by MHC molecules at their surface. Thus, killing is provoked by direct recognition of a target on the β -cell and requires T-cell/ β -cell contact. In mice, this would imply recognition of MHC class-I-restricted antigens by $CD8^+$ T cells because their β -cells do not express MHC class II molecules. Several observations are consistent with this scheme: (1) $CD8^+$ T cells follow $CD4^+$ T cells into the islets when splenocytes from diabetic donors are transferred into healthy NOD recipients; (2) both $CD4^+$ and $CD8^+$ T cells are required for transfer of diabetes even though the former can invade the host's islets alone; (3) some $CD8^+$ T-cell clones derived from diabetic NOD mice can provoke diabetes on transfer into lymphocyte-deficient recipients, or are diabetogenic in monoclonal TCR transgenic mice^{33,34}; (4) diabetes but not insulinitis is reduced when insulinitic, pre-diabetic NOD mice are injected with anti-MHC class I or anti- $CD8$ monoclonal antibodies; (5) $CD8^+$ T cells isolated from diabetic animals can attach to and lyse β -cells *in vitro*. Also perhaps relevant is that $CD8^+$ T cells frequently dominate islet infiltrates in humans with diabetes (for example, see ref. 35). However, other evidence argues against this scheme: (1) some $CD4^+$ T-cell clones can provoke diabetes in the absence of $CD8^+$ T cells in transfer or TCR transgenic systems³⁶; (2) splenocytes from diabetic donors can transfer disease into recipients lacking MHC class I molecules on their islets³⁷; and (3) diabetes can develop in a TCR transgenic model in the absence of antigen-specific contact between T and β -cells³⁸.

The second, 'activation-linked', mechanism is an indirect one, by which the mere proximity of β -cells to angry T cells provokes their death. This scheme is more consistent with two perplexing observations: $CD4^+$ T cells can cause diabetes in mice in the absence of $CD8^+$ T cells; but MHC class II molecules are not detected on murine β -cells. According to this scheme, β -cell autoantigens are picked up by APCs (macrophages or DCs) resident in or recruited to the islets, and are presented to primed $CD4^+$ T cells provoked to circulate through the tissues. As a result of such T-cell/APC interactions, the activated T cells may directly kill the bystander β -cells, for example through perforin or Fas/Fas ligand (FasL) interactions (pathway i in Fig. 3b); produce soluble mediators that induce β -cell death (ii in Fig. 3b); or activate the cytotoxic functions of macrophages in the vicinity (iii in Fig. 3b). The β -cells themselves may enter the melee, induced to synthesize soluble mediators that hasten their own demise (iv in Fig. 3b). The nature of the soluble mediators involved is controversial: interferon (IFN)- γ , IL-1, TNF- α , IL-6 and nitric oxide (NO) have all been implicated (see below).

Molecular mechanisms: cell-cell contact

Recent efforts have been directed at dissecting the molecular mechanisms responsible for the β -cell death that preludes type 1 diabetes³⁹. According to several mouse models^{36,40}, apoptosis is probably the main mode of destruction. For humans, finer mechanistic studies are essentially restricted to experiments on β -cells cultured *in vitro* or expression profiling of *ex vivo* material; for mouse models, the range of genetically engineered knockout strains has added an important perspective. Nonetheless, the results acquired so far have often been contradictory and unsatisfying.

In recognition-linked killing (and also the surface-receptor-mediated variant of activation-linked killing; Fig. 3b, pathway i), most attention has been devoted to perforin and Fas/FasL. The perforin/granzyme pathway (Fig. 2), which leads to insertion of tubular perforin complexes into the cell membrane and consequent osmotic lysis, is the main mode of death induced by cytotoxic T cells.

Both NOD mice^{41,42} and a class-I-restricted TCR transgenic model⁴³ showed normal insulinitis, but a reduced incidence and delayed onset of diabetes when rendered perforin-deficient, suggesting that perforin has a role in terminal β -cell destruction. In direct contradiction to these results, however, two class-I-restricted TCR transgenic models developed severe diabetes on a perforin-deficient background^{42,44}, and a third model also did neonatally, although disease was attenuated when perforin-deficient transgenic T cells were transferred into a normal recipient⁴⁵. Perhaps the explanation for these inconsistent data lies in other effects of the perforin mutation; for example, perforin also seems to be involved in the homeostasis of memory $CD8^+$ cells⁴⁶. Alternatively, different $CD8^+$ T-cell clones, recognizing various β -cell antigens, might intervene at diverse stages in the disease process, relying on different effector mechanisms.

Results on the role of Fas/FasL in β -cell death have been equally as confusing. Fas, which is expressed on activated lymphocytes and some cells in tissues, is probably the main cell-surface receptor transmitting death signals (Fig. 2), and FasL is expressed primarily on activated T lymphocytes⁴⁷. The first evidence that this receptor/ligand pair might be involved in terminal β -cell destruction came from the observation that NOD mice homozygous for the *lpr* mutation (a mutation of the *fas* gene) did not develop diabetes; moreover, transfer of splenocytes from diabetic standard NOD mice into NOD *lpr/lpr* recipients did not provoke diabetes, although it did in wild-type hosts⁴⁸. The homozygous *lpr* mutation also had an effect in one TCR transgenic model⁴², but not in another⁴⁵.

These findings prompted several studies demonstrating Fas on β -cells isolated from rodents and humans, and its upregulation during progression to diabetes (for example, see ref. 48), although the

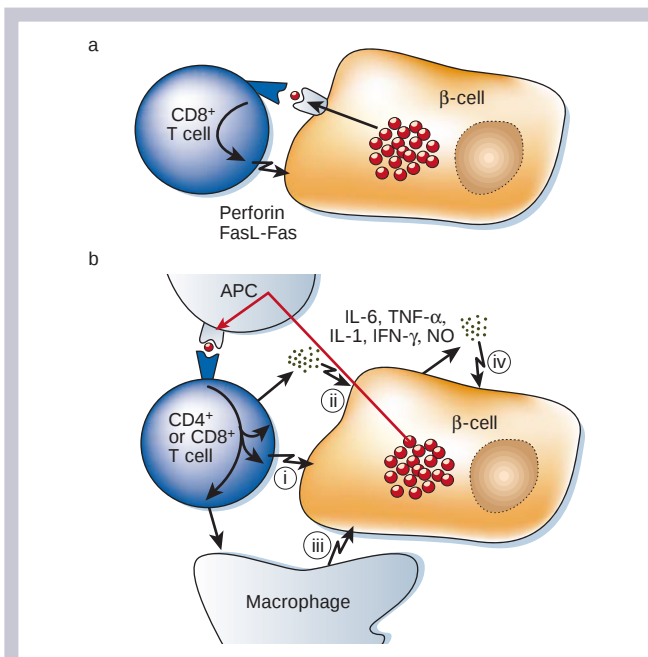


Figure 3 Proposed cellular mechanisms of β -cell death. **a**, According to 'recognition-linked' mechanisms, the T cell (in this case $CD8^+$ T cell) is activated by direct recognition of islet β -cell antigens (red dots) presented by MHC molecules (in this case class I molecules) on β -cells. Activation provokes killing of the β -cell through cell/cell contact using, for example, the Fas/FasL or perforin pathways. **b**, According to 'activation-linked' mechanisms, the T cell (either $CD4^+$ or $CD8^+$) recognizes β -cell antigens presented indirectly by APCs (macrophages, DCs) in the vicinity. The resulting activation provokes β -cell death mediated by surface receptors (such as FasL/Fas, TNF- α /TNF-R) (i); cytokines and other soluble death mediators produced by the T cell (ii); activation of macrophages and stimulation of their cytotoxic activities (iii); and activation of β -cells and stimulation of their production of cell death mediators (iv).

complex cellular composition of inflamed islets renders histological interpretation difficult, and Fas upregulation and its function have been disputed⁴⁹. In addition, *in vitro* studies have documented cytokine-induced Fas upregulation on cultured murine and human islets and its ability to transmit apoptosis signals^{49,50}. But it has also been reported that insulinitis was abrogated in NOD *lpr/lpr* mice⁵¹, meaning that the lack of diabetes in these animals might have nothing to do with terminal β -cell destruction, but rather to another pleiotropic effect of this mutation on the lymphoid system or to a need for some β -cell death to initiate the autoimmune response (see above).

Indeed, islets from NOD *lpr/lpr* mice were not protected from autoimmune attack when transferred into diabetic wild-type recipients^{52,53}, and even wild-type islets were protected when introduced into NOD *lpr/lpr* hosts⁵². The explanation for these contradictory transfer results seems to be that lymphocytes residing in *lpr/lpr* mice strongly upregulate FasL⁵⁴, rendering them potent killers of any introduced cells expressing Fas⁵⁵. This neat explanation is not completely satisfying, however, as it has been reported that a homozygous *gld* (or FasL) mutation, which lacks the pleiotropic effects of the *lpr* mutation, also inhibited diabetes in NOD mice, and that transfer of diabetogenic T cells into lymphocyte-deficient NOD severe combined immunodeficient (SCID) *lpr/lpr* mice also resulted in attenuation of diabetes⁵⁶.

A role for membrane TNF/TNF receptor II (TNF-RII) interactions in recognition-linked killing of β -cells in a CD8⁺ T-cell-dependent diabetes model has been suggested⁵⁷, although, again, these data are difficult to interpret because of the pleiotropic mutations in the mouse strains from which they derive. Interestingly, however, TNF-RII expression was upregulated during disease unfolding in a CD4⁺ T-cell-dependent TCR transgenic model⁵⁸.

Molecular mechanisms: soluble mediators

An implication of perforin, Fas/FasL or membrane TNF- α /TNF-R in terminal β -cell destruction cannot be taken as evidence for a recognition-linked mode of death, because T-cell/ β -cell contact is also a feature of a variant of the activation-linked mechanism (pathway i in Fig. 3b). Nonetheless, the other variants all entail soluble mediators (pathways ii–iv in Fig. 3b) and, as discussed above, because there is evidence to support this latter class of mechanisms, much effort has been devoted to identifying the factors that might be involved.

Untangling the relative roles of the different pro-inflammatory cytokines in β -cell death is problematic because these cytokines are produced by more than one cell type and have effects on many targets. In addition, the same cytokine might act as a membrane or soluble form (with a different range of activities) or might engage several receptors (again with different effects). Moreover, the pro-inflammatory cytokines form a network of molecules, which influence each other's synthesis and activities. The difficulties encountered are illustrated by TNF- α . This cytokine exists in a soluble or membrane-bound form, interacts with TNF-RI and -RII, and promotes activities, such as inflammation, differentiation, survival and death, in many cell types⁵⁹. Elevated levels of TNF- α expression have been correlated with β -cell destruction *in vivo*⁶⁰, and TNF- α can contribute directly or indirectly to β -cell destruction *in vitro*^{61,62}, but its effects on the unfolding of diabetes *in vivo* are confusing. Systemic injection of TNF- α gave different results according to the age of the recipient: enhancing disease before 3 weeks, and blocking afterwards⁶³. This differential effect has been mimicked in transgenic mice expressing TNF- α specifically in the β -cells of NOD mice⁴. These results probably reflect changing populations of TNF- α producers and users throughout the development of the animal and disease progression.

Thus, interpreting experiments using mice with a genetic deficiency in TNF- α may not be straightforward. NOD mice deficient in TNF-RI developed normal insulinitis but reduced diabetes; they also showed less than the usual diabetes when transferred with T cells from wild-type NOD donors⁶⁴. To circumvent some of the problems,

a transplant model was developed with the transfer of both islets from various knockout mice and diabetogenic lymphocytes from TCR transgenic mice into NOD-SCID recipients. Transplants carrying mutations in TNF-RI, but not Fas, IFN- γ receptor or inducible NO synthetase (iNOS) resisted rejection, although this turned out to reflect an indirect effect on lymphocyte residence in the islets rather than a direct effect on β -cell killing⁶⁵.

Another difficult area is the role of free radicals and reactive oxygen intermediates in terminal β -cell destruction³⁹. Inflammation leads to the induction of iNOS and consequent production of NO in macrophages and other cell types. IL-1, TNF- α and IFN- γ have all been implicated in this increased production, but IL-1 β may be primordial⁶⁶. Elevated concentrations of NO lead to DNA damage and apoptosis by a poorly defined mechanism. More specifically related to diabetes, *in vitro* both macrophages⁶⁷ and islet β -cells⁶⁸ can be induced to synthesize NO, and both exogenously and endogenously produced NO provokes β -cell destruction⁶¹. *In vivo*, iNOS concentrations increase in parallel with diabetes onset in a model of T-cell transfer⁶⁹ and after cyclophosphamide treatment of insulinitic animals⁷⁰; moreover, an inhibitor of iNOS activity can block diabetes transfer⁶⁹. Thus, it was surprising to find that iNOS-deficient NOD mice developed diabetes as usual (J. Fenyk-Melody and J. Mudger, personal communication) and that islets from such animals were attacked efficiently by diabetogenic T cells in the transplant model mentioned above⁶⁵.

Currently, debate is focused on just how critical NO is in 'natural' diabetes, and whether the main culprit is NO produced exogenously by macrophages or NO produced endogenously by β -cells (or both). Resolving the latter issue is important because it might suggest a mechanism for the specificity of β -cell death: at least *in vitro*, both islet α - and β -cells are sensitive to NO-induced destruction⁶⁶, but only the latter show elevated iNOS concentrations after cytokine treatment⁶⁶. Notably, cytokine-induced death of human β -cells seems not to depend critically on iNOS *in vitro*⁷¹, but it has been suggested that NO serves to induce the expression of Fas on human islets, rendering them susceptible to T-cell attack⁷².

Other oxygen-reactive species might also be involved in β -cell destruction during diabetes development. Peroxynitrite⁷³ and hydrogen peroxide⁷⁴ are involved in β -cell killing *in vitro*, whereas overexpression of catalase⁷⁵, manganese superoxide dismutase^{75,76}, glutathione peroxidase⁷⁵ and thioredoxin⁷⁷ have such an influence on cytokine-mediated killing *in vitro* or on diabetes development in NOD mice. In addition, the ALR mouse strain constitutively expresses, both systemically and in β -cells, high levels of molecules associated with dissipating free-radical stress, including manganese superoxide dismutase and glutathione peroxidase, and islets from these animals resist cytokine-induced β -cell death through cytokines *in vitro* and through diabetogenic T cells *in vitro* and *in vivo*⁷⁸.

Why is it so complex?

So the picture that we currently have is blurred: data from different diabetes models implicate distinct death effector systems with no clear indication of which, if any, are involved in the development of human diabetes. It should be remembered that different intracellular death signalling pathways may be implicated in the β -cell apoptosis induced by diverse agents⁷⁹, and that murine and human β -cells may show variable susceptibilities to apoptosis mediated by the different pathways⁸⁰. In particular, one has to question the relevance of *in vitro* studies in light of increasing evidence that the removal of cells from their normal tissue environment can launch a very specific cell death program, 'anoikis', mediated by *bmf* — a member of the pro-apoptotic Bcl-2 family⁸¹.

It is also important to remember that affected individuals can present with heterogeneous clinical parameters⁸², and thus diabetes in man could be a set of related disorders with possible differences in inciting antigen, primary effector cell type and, most relevant here, mechanism of β -cell destruction. It is not known which of the mouse

models will prove best for studying each of the human variants. Neither is it known whether islet grafts are subject to the same or different mechanisms of destruction — an important issue given the increasing interest in using grafts to attenuate disease.

Death as a connector

By contrast, type 2 diabetes is non-autoimmune in nature⁸³. Its hallmarks are insulin resistance issuing from any of several causes and inadequate insulin secretion. In advanced stages of disease, β -cell function may degenerate to such a degree that insulin therapy becomes necessary. As mentioned above, a significant fraction of individuals originally diagnosed with type 2 diabetes are cryptic type 1 diabetics or evolve with time to a type 1 state, which is defined as exhibiting anti- β -cell autoimmunity. What connects the two types?

The nature of the β -cell abnormality underlying reduced insulin secretion in humans with late type 2 diabetes may be multifactorial. Hyperglycaemia can lead to alterations in glucose sensing by the β -cell, and evidence has been presented for a genetic component that results in defects in β -cell function, a reduction in β -cell mass, or a combination of the two. A modest reduction in the β -cell mass of type 2 individuals at autopsy has been reported⁸⁴, but conflicting results abound^{85,86}. When β -cell mass is reduced, amyloid deposits are often seen^{84,86}, but the extent to which they contribute to β -cell failure has been disputed⁸⁵. It may be worth keeping in mind that type 2 diabetics would be expected to have an elevated β -cell mass owing to hyperplasia, so the comparable masses sometimes found in diabetic compared with normal individuals might still reflect induction of β -cell death — only longitudinal studies can give a true indication. *In vitro* experiments have shown that elevated glucose concentrations, analogous to those seen in type 2 diabetics, can induce apoptosis of cultured human islets⁸⁷.

Better evidence for β -cell death in late stages of type 2 diabetes comes from animal models. Perhaps the most convincing are data from the gerbil, *Psammomys obesus*, which does not naturally develop diabetes, but can in captivity when fed a high-energy diet. The disease begins with a state of insulin resistance/hyperinsulinaemia/normoglycaemia and evolves to one of hypoinsulinaemia/hyperglycaemia. During the development of hyperglycaemia, an increase in apoptosis occurs, culminating in a severe disruption of the islet architecture^{88,89}. This increased β -cell death could result from hyperglycaemia, as suggested by *in vitro* studies showing that islets from diabetes-prone gerbils exhibit a progressive apoptosis on culture at high glucose concentrations^{89,90}. *In vitro* studies on other rodent models^{91,92}, have also shown apoptosis of β -cells after incubation in elevated concentrations of glucose or free fatty acids.

Almost nothing is known about the mechanism(s) of this β -cell death. One possible clue has come from studies on islets from human type 2 diabetics at autopsy showing a decrease in the ratio of anti-apoptotic to pro-apoptotic members of the Bcl-2 family: decreased Bcl-x_i; increased Bad, Bid and Bik⁸⁷. Interestingly, increased levels of apoptosis and islet cell Bad were also characteristic of mice that are diabetic because of a genetic deficiency in insulin receptor substrates (IRS-2)⁹³. These findings support the notion that apoptosis signals channel through the mitochondria and argue against surface-receptor-mediated signals (Fig. 2).

Another clue may come from Perk knockout mice, which develop diabetes at an early age owing to massive disappearance of β -cells⁹⁴. The protein kinase Perk couples protein folding in the ER to polypeptide biosynthesis by phosphorylating one of the subunits of eukaryotic translation initiation factor 2 (eIF2 α), which attenuates translation in response to ER stress such as insulin biosynthesis in states of hyperglycaemia or hyperlipidaemia^{94,95}. In the absence of Perk, translation could not be muted by this mechanism, the ER became distended, the ER stress transducer IRE1 α was activated, and massive loss of β -cells and diabetes occurred. IRE1 α lies upstream of caspase-12 (ref. 28), which is specifically activated during the ER stress response (ref. 26; and Fig. 2), so it is likely that this is the death

effector pathway mobilized in such contexts. Notably, individuals affected with Wolcott–Rallison syndrome develop diabetes as infants and have a mutation in the human equivalent of Perk⁹⁶. Thus, abnormalities in Perk or associated molecules might underlie any reduced β -cell mass that occurs in the more usual cases of type 2 diabetes.

It seems, then, that β -cell death may be what connects type 2 and type 1 diabetes in those occasional cases where the former evolves into the latter. Depending on genetic and/or environmental influences on a particular individual's course of pathogenesis, a fraction of type 2 diabetics may exhibit abnormal β -cell death; and, again reflecting a combination of genetic and environmental effects, a subfraction of these may show mobilization of T cells reactive to β -cell autoantigens, eventually culminating in autoimmune, insulin-dependent diabetes. If so, the scheme hypothesized to unfold is very reminiscent of the one that we proposed above for initiating type 1 diabetes in more standard cases, as are some of the players speculated to take part (metabolic fluxes, ER stress, caspase-12). Animal models permitting direct comparison of early- and late-onset autoimmune diabetes need to be engineered to evaluate this notion. □

1. Tisch, R. & McDevitt, H. Insulin-dependent diabetes mellitus. *Cell* **85**, 291–297 (1996).
2. Bach, J. F. & Mathis, D. 70th forum in immunology: the NOD mouse. *Res. Immunol.* **148**, 281–370 (1997).
3. Hoglund, P. *et al.* Initiation of autoimmune diabetes by developmentally regulated presentation of islet cell antigens in the pancreatic lymph nodes. *J. Exp. Med.* **189**, 331–339 (1999).
4. Green, E. A., Eynon, E. E. & Flavell, R. A. Local expression of TNF α in neonatal NOD mice promotes diabetes by enhancing presentation of islet antigens. *Immunity* **9**, 733–743 (1998).
5. Katz, J. D., Benoist, C. & Mathis, D. T helper cell subsets in insulin-dependent diabetes. *Science* **268**, 1185–1188 (1995).
6. Gonzalez, A. *et al.* Genetic control of diabetes progression. *Immunity* **7**, 873–883 (1997).
7. Finegood, D. T., Scaglia, L. & Bonner-Weir, S. Dynamics of β -cell mass in the growing rat pancreas. *Diabetes* **44**, 249–256 (1995).
8. Scaglia, L., Cahill, C. J., Finegood, D. T. & Bonner-Weir, S. Apoptosis participates in the remodeling of the endocrine pancreas in the neonatal rat. *Endocrinology* **138**, 1736–1741 (1997).
9. Petrik, J., Arany, E., McDonald, T. J. & Hill, D. J. Apoptosis in the pancreatic islet cells of the neonatal rat is associated with a reduced expression of insulin-like growth factor II that may act as a survival factor. *Endocrinology* **139**, 2994–3004 (1998).
10. Trudeau, J. D. *et al.* Neonatal β -cell apoptosis: a trigger for autoimmune diabetes? *Diabetes* **49**, 1–7 (2000).
11. Kassem, S. A., Ariel, I., Thornton, P. S., Scheimberg, I. & Glaser, B. β -cell proliferation and apoptosis in the developing normal human pancreas and in hyperinsulinism of infancy. *Diabetes* **49**, 1325–1333 (2000).
12. Steinman, R. M., Turley, S., Mellman, I. & Inaba, K. The induction of tolerance by dendritic cells that have captured apoptotic cells. *J. Exp. Med.* **191**, 411–416 (2000).
13. Bellone, M. *et al.* Processing of engulfed apoptotic bodies yields T cell epitopes. *J. Immunol.* **159**, 5391–5399 (1997).
14. Pittoni, V. & Isenberg, D. Apoptosis and antiphospholipid antibodies. *Semin. Arthritis Rheum.* **28**, 163–178 (1998).
15. Coles, H. S. R., Burne, J. F. & Raff, M. C. Large-scale normal cell death in the developing rat kidney and its reduction by epidermal growth factor. *Development* **118**, 777–784 (1993).
16. Meier, P., Finch, A. & Evan, G. Apoptosis in development. *Nature* **407**, 796–801 (2001).
17. Chautan, M., Chazal, G., Cecconi, F., Gruss, P. & Golstein, P. Interdigital cell death can occur through a necrotic and caspase-independent pathway. *Curr. Biol.* **9**, 967–970 (1999).
18. Jacobson, M. D., Weil, M. & Raff, M. C. Programmed cell death in animal development. *Cell* **88**, 347–354 (1997).
19. Mayor, F. & Cuevas, J. M. Hormonal and metabolic changes in the perinatal period. *Biol. Neonate* **48**, 185–196 (1985).
20. Parham, P. Intolerable secretion in tolerant transgenic mice. *Nature* **333**, 500–503 (1988).
21. Li, K. *et al.* Cytochrome c deficiency causes embryonic lethality and attenuates stress-induced apoptosis. *Cell* **101**, 389–399 (2000).
22. Cecconi, F., Alvarez-Bolado, G., Meyer, B. I., Roth, K. A. & Gruss, P. Apaf1 (CED-4 homolog) regulates programmed cell death in mammalian development. *Cell* **94**, 727–737 (1998).
23. Yoshida, H. *et al.* Apaf1 is required for mitochondrial pathways of apoptosis and brain development. *Cell* **94**, 739–750 (1998).
24. Hakem, R. *et al.* Differential requirement for caspase 9 in apoptotic pathways in vivo. *Cell* **94**, 339–352 (1998).
25. Kuida, K. *et al.* Reduced apoptosis and cytochrome c-mediated caspase activation in mice lacking caspase 9. *Cell* **94**, 325–337 (1998).
26. Nakagawa, T. *et al.* Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid- β . *Nature* **403**, 98–103 (2000).
27. Rao, R. V. *et al.* Coupling endoplasmic reticulum stress to the cell death program: mechanism of caspase activation. *J. Biol. Chem.* (in the press).
28. Nakagawa, T. & Yuan, J. Cross-talk between two cysteine protease families. Activation of caspase 12 by calpain in apoptosis. *J. Cell Biol.* **150**, 887–894 (2000).
29. Yoneda, T. *et al.* Activation of caspase-12, an endoplasmic reticulum (ER) resident caspase, through tumor necrosis factor receptor-associated factor 2-dependent mechanism in response to the ER stress. *J. Biol. Chem.* **276**, 13935–13940 (2001).
30. Wegmann, D. R. The immune response to islets in experimental diabetes and insulin-dependent diabetes mellitus. *Curr. Opin. Immunol.* **8**, 860–864 (1996).

31. Benoist, C. & Mathis, D. Autoimmunity provoked by infection: how good is the case for T cell epitope mimicry? *Nature Immunol.* **2**, 797–801 (2001).
32. Huang, F.-B. *et al.* A discrete subpopulation of dendritic cells transports apoptotic intestinal epithelial cells to T cell areas of mesenteric lymph nodes. *J. Exp. Med.* **191**, 435–443 (2000).
33. Wong, F. S., Visintin, L., Wen, L., Flavell, R. A. & Janeway, C. A. CD8 T cell clones from young nonobese diabetic (NOD) islets can transfer rapid onset of diabetes in NOD mice in the absence of CD4 cells. *J. Exp. Med.* **183**, 67–76 (1996).
34. Graser, R. T. *et al.* Identification of a CD8 T cell that can independently mediate autoimmune diabetes development in the complete absence of CD4 T cell helper functions. *J. Immunol.* **164**, 3913–3918 (2000).
35. Itoh, N. *et al.* Mononuclear cell infiltration and its relation to the expression of major histocompatibility complex antigens and adhesion molecules in pancreas biopsy specimens from newly diagnosed insulin-dependent diabetes mellitus patients. *J. Clin. Invest.* **92**, 2313–2322 (1993).
36. Kurrer, M. O., Pakala, S. V., Hanson, H. L. & Katz, J. D. β cell apoptosis in T cell-mediated autoimmune diabetes. *Proc. Natl Acad. Sci. USA* **94**, 213–218 (1997).
37. Serreze, D. V. *et al.* Initiation of autoimmune diabetes in NOD/Lt mice is MHC class I-dependent. *J. Immunol.* **158**, 3978–3986 (1997).
38. Sarukhan, A., Lechner, O. & von Boehmer, H. Autoimmune insulinitis and diabetes in the absence of antigen-specific contact between T cells and islet β -cells. *Eur. J. Immunol.* **29**, 3410–3416 (1999).
39. Thomas, H. E. & Kay, T. W. H. Beta cell destruction in the development of autoimmune diabetes in the non-obese diabetic (NOD) mouse. *Diabetes Metab. Res. Rev.* **16**, 251–261 (2000).
40. O'Brien, B. A., Harmon, B. V., Cameron, D. P. & Allan, D. J. Apoptosis is the mode of beta-cell death responsible for the development of IDDM in the nonobese diabetic (NOD) mouse. *Diabetes* **46**, 750–757 (1997).
41. Kagi, D. *et al.* Reduced incidence and delayed onset of diabetes in perforin-deficient nonobese diabetic mice. *J. Exp. Med.* **186**, 989–997 (1997).
42. Amrani, A. *et al.* Perforin-independent β -cell destruction by diabetogenic CD8⁺ T lymphocytes in transgenic nonobese diabetic mice. *J. Clin. Invest.* **103**, 1201–1209 (1999).
43. Kagi, D., Odermatt, B., Ohashi, P. S., Zinkernagel, R. M. & Hengartner, H. Development of insulinitis without diabetes in transgenic mice lacking perforin-dependent cytotoxicity. *J. Exp. Med.* **183**, 2143–2152 (1996).
44. Seewaldt, S. *et al.* Virus-induced autoimmune diabetes: most beta-cells die through inflammatory cytokines and not perforin from autoreactive (anti-viral) cytotoxic T-lymphocytes. *Diabetes* **49**, 1801–1809 (2000).
45. Kreuwel, H. T. *et al.* Comparing the relative role of perforin/granzyme versus Fas/Fas ligand cytotoxic pathways in CD8⁺ T cell-mediated insulin-dependent diabetes mellitus. *J. Immunol.* **163**, 4335–4341 (1999).
46. Badovinac, V. P., Tivnnerheim, A. R. & Harty, J. T. Regulation of antigen-specific CD8⁺ T cell homeostasis by perforin and interferon- γ . *Science* **290**, 1354–1358 (2000).
47. Krammer, P. H. CD95's deadly mission in the immune system. *Nature* **407**, 789–795 (2000).
48. Chervonsky, A. V. *et al.* The role of Fas in autoimmune diabetes. *Cell* **89**, 17–24 (1997).
49. Thomas, H. E., Darwiche, R., Corbett, J. A. & Kay, T. W. H. Evidence that β cell death in the nonobese diabetic mouse is Fas independent. *J. Immunol.* **163**, 1562–1569 (1999).
50. Loweth, A. C., Williams, G. T., James, R. F. L., Scarpello, J. H. B. & Morgan, N. G. Human islets of Langerhans express Fas ligand and undergo apoptosis in response to interleukin-1 β and Fas ligation. *Diabetes* **47**, 727–732 (1998).
51. Itoh, N. *et al.* Requirement of Fas for the development of autoimmune diabetes in nonobese diabetic mice. *J. Exp. Med.* **186**, 613–618 (1997).
52. Allison, J. & Strasser, A. Mechanisms of beta cell death in diabetes: a minor role for CD95. *Proc. Natl Acad. Sci. USA* **95**, 13818–13822 (1998).
53. Kim, Y. H. *et al.* Apoptosis of pancreatic beta-cells detected in accelerated diabetes of NOD mice: no role of Fas-Fas ligand interaction in autoimmune diabetes. *Eur. J. Immunol.* **29**, 455–465 (1999).
54. Watanabe, D., Suda, T., Hashimoto, H. & Nagata, S. Constitutive activation of the Fas ligand gene in mouse lymphoproliferative disorders. *EMBO J.* **14**, 12–18 (1995).
55. Kim, S. *et al.* Inhibition of autoimmune diabetes by Fas ligand: the paradox is solved. *J. Immunol.* **164**, 2931–2936 (2000).
56. Su, X. *et al.* Significant role for Fas in the pathogenesis of autoimmune diabetes. *J. Immunol.* **164**, 2523–2532 (2000).
57. Herrera, P. L., Harlan, D. M. & Vassalli, P. A mouse CD8 T cell-mediated acute autoimmune diabetes independent of the perforin and Fas cytotoxic pathways: possible role of membrane TNF. *Proc. Natl Acad. Sci. USA* **97**, 279–284 (2000).
58. Walter, U. *et al.* Monitoring gene expression of TNFR family members by β -cells during development of autoimmune diabetes. *Eur. J. Immunol.* **30**, 1224–1232 (2000).
59. Beutler, B. A. The role of tumor necrosis factor in health and disease. *J. Rheumatol.* **57**, 16–21 (1999).
60. Mueller, C., Held, W., Imboden, M. A. & Carnaud, C. Accelerated β -cell destruction in adoptively transferred autoimmune diabetes correlates with an increased expression of the genes coding for TNF- α and granzyme A in the intra-islet infiltrates. *Diabetes* **44**, 112–117 (1995).
61. Kaneto, H. *et al.* Apoptotic cell death triggered by nitric oxide in pancreatic β -cells. *Diabetes* **44**, 733–738 (1995).
62. Mandrup-Poulsen, T., Bendtzen, K., Dinarello, C. A. & Nerup, J. Human tumor necrosis factor potentiates human interleukin 1-mediated rat pancreatic beta-cell cytotoxicity. *J. Immunol.* **139**, 4077–4082 (1987).
63. Yang, X.-D. *et al.* Effect of tumor necrosis factor α on insulin-dependent diabetes mellitus in NOD mice. I. The early development of autoimmunity and the diabetogenic process. *J. Exp. Med.* **180**, 995–1004 (1994).
64. Kagi, D. *et al.* TNF receptor 1-dependent β cell toxicity as an effector pathway in autoimmune diabetes. *J. Immunol.* **162**, 4598–4605 (1999).
65. Pakala, S. V., Chivetta, M., Kelly, C. B. & Katz, J. D. In autoimmune diabetes the transition from benign to pernicious insulinitis requires an islet cell response to tumor necrosis factor α . *J. Exp. Med.* **189**, 1053–1062 (1999).
66. Corbett, J. A. & McDaniel, M. L. Intra-islet release of interleukin 1 inhibits β cell function by inducing β cell expression of inducible nitric oxide synthase. *J. Exp. Med.* **181**, 559–568 (1995).
67. Kroncke, K. D., Funda, J., Berschick, B., Kolb, H. & Kolb-Bachofen, V. Macrophage cytotoxicity towards isolated rat islet cells: neither lysis nor its protection by nicotinamide are beta-cell specific. *Diabetologia* **34**, 232–238 (1991).
68. Corbett, J. A., Lancaster, J. R. Jr, Sweetland, M. A. & McDaniel, M. L. Interleukin-1 β -induced formation of EPR-detectable iron-nitrosyl complexes in islets of Langerhans. *J. Biol. Chem.* **266**, 21351–21354 (1991).
69. Corbett, J. A. *et al.* Nitric oxide production in islets from nonobese diabetic mice: aminoguanidine-sensitive and -resistant stages in the immunological diabetic process. *Proc. Natl Acad. Sci. USA* **90**, 8992–8995 (1993).
70. Rothe, H., Burkart, V., Faust, A. & Kolb, H. Interleukin-12 gene expression is associated with rapid development of diabetes mellitus in non-obese diabetic mice. *Diabetologia* **39**, 119–122 (1996).
71. Rabinovitch, A. *et al.* Human pancreatic islet beta-cell destruction by cytokines is independent of nitric oxide production. *J. Clin. Endocrinol. Metab.* **79**, 1058–1062 (1994).
72. Stassi, G. *et al.* Nitric oxide primes pancreatic beta cells for Fas-mediated destruction in insulin-dependent diabetes mellitus. *J. Exp. Med.* **186**, 1193–1200 (1997).
73. Delaney, C. A. *et al.* Sensitivity of human pancreatic islets to peroxynitrite-induced cell dysfunction and death. *FEBS Lett.* **394**, 300–306 (1996).
74. Rabinovitch, A., Suarez, W. L., Thomas, P. D., Strynadka, K. & Simpson, I. Cytotoxic effects of cytokines on rat islets: evidence for involvement of free radicals and lipid peroxidation. *Diabetologia* **35**, 409–413 (1992).
75. Lortz, S. *et al.* Protection of insulin-producing RINm5F cells against cytokine-mediated toxicity through overexpression of antioxidant enzymes. *Diabetes* **49**, 1123–1130 (2000).
76. Hohmeier, H. E., Thigpen, A., Tran, W., Davis, R. & Newgard, C. B. Stable expression of manganese superoxide dismutase (MnSOD) in insulinoma cells prevents IL-1 β -induced cytotoxicity and reduces nitric oxide production. *J. Clin. Invest.* **101**, 1811–1820 (1998).
77. Hotta, M. *et al.* Pancreatic β cell-specific expression of thioredoxin, an antioxidant and antiapoptotic protein, prevents autoimmune and streptozotocin-induced diabetes. *J. Exp. Med.* **188**, 1445–1451 (1998).
78. Mathews, C. E., Graser, R. T., Savinov, A., Serreze, D. V. & Leithner, E. H. Unusual resistance of ALR/Lt mouse beta cells to autoimmune destruction: role for beta cell-expressed resistance determinants. *Proc. Natl Acad. Sci. USA* **98**, 235–240 (2001).
79. Allison, J. *et al.* Transgenic overexpression of human Bcl-2 in islet β cells inhibits apoptosis but does not prevent autoimmune destruction. *Int. Immunol.* **12**, 9–17 (2000).
80. Eizirik, D. L. *et al.* Major species differences between humans and rodents in the susceptibility to pancreatic β -cell injury. *Proc. Natl Acad. Sci. USA* **91**, 9253–9256 (1994).
81. Klaus, G. G., Pepys, M. B., Kitajima, K. & Askonas, B. A. Activation of mouse complement by different classes of mouse antibody. *Immunology* **38**, 687–695 (1979).
82. Bach, J. F. Insulin-dependent diabetes mellitus as an autoimmune disease. *Endocr. Rev.* **15**, 516–542 (1994).
83. Taylor, S. I. Deconstructing type 2 diabetes. *Cell* **97**, 9–12 (1999).
84. Shimabukuro, M., Ohneda, M., Lee, Y. & Unger, R. H. Islet amyloid, increased A-cells, reduced B-cells and exocrine fibrosis: quantitative changes in the pancreas in type 2 diabetes. *Diabetes Res.* **9**, 151–159 (1988).
85. Sempoux, C., Guio, Y., Dubois, D., Moulin, P. & Rahier, J. Human type 2 diabetes: morphological evidence for abnormal β -cell function. *Diabetes* **50**, S172–S177 (2001).
86. Guio, Y., Sempoux, C., Moulin, P. & Rahier, J. No decrease of the β -cell mass in type 2 diabetic patients. *Diabetes* **50**, S188 (2001).
87. Federici, M. *et al.* High glucose causes apoptosis in cultured human pancreatic islets of Langerhans. *Diabetes* **50**, 1290–1301 (2001).
88. Bar-On, H., Ben-Sasson, R., Ziv, E., Arar, N. & Shafir, E. Irreversibility of nutritionally induced NIDDM in *Psammomys obesus* is related to beta-cell apoptosis. *Pancreas* **18**, 259–265 (1999).
89. Donath, M. Y., Gorss, D. J., Cerasi, E. & Kaiser, N. Hyperglycemia-induced β -cell apoptosis in pancreatic islets of *psammomys obesus* during development of diabetes. *Diabetes* **48**, 738–744 (1999).
90. Leibowitz, G. *et al.* β -cell glucotoxicity in the *psammomys obesus* model of type 2 diabetes. *Diabetes* **50**, S113–S117 (2001).
91. Shimabukuro, M., Zhou, Y. T., Levi, M. & Unger, R. H. Fatty acid-induced beta cell apoptosis: a link between obesity and diabetes. *Proc. Natl Acad. Sci. USA* **95**, 2498–2502 (1998).
92. Efanova, I. B. *et al.* Glucose and tolbutamide induce apoptosis in pancreatic beta-cells: a process dependent on intracellular Ca^{2+} concentration. *J. Biol. Chem.* **273**, 33501–33507 (1998).
93. Withers, D. J. *et al.* Irs-2 coordinates Igf-1 receptor-mediated β -cell development and peripheral insulin signalling. *Nature Genet.* **23**, 32–40 (1999).
94. Harding, H. P. *et al.* Diabetes mellitus and exocrine pancreatic dysfunction in *perk*^{-/-} mice reveals a role for translational control in secretory cell survival. *Mol. Cell* **7**, 1153–1163 (2001).
95. Gilligan, M. *et al.* Glucose stimulates the activity of the guanine nucleotide-exchange factor eIF-2B in isolated rat islets of Langerhans. *J. Biol. Chem.* **271**, 2121–2125 (1996).
96. Delepine, M. *et al.* EIF2AK3, encoding translation initiation factor 2- α kinase 3, is mutated in patients with Wolcott-Rallison syndrome. *Nature Genet.* **25**, 406–409 (2000).
97. Guerdier, S., Eynon, E. E. & Flavell, R. A. Autoimmunity without diabetes in transgenic mice expressing β cell-specific CD86, but not CD80: parameters that trigger progression to diabetes. *J. Immunol.* **161**, 2128–2140 (1998).
98. Katz, J. D., Wang, B., Haskins, K., Benoist, C. & Mathis, D. Following a diabetogenic T cell from genesis through pathogenesis. *Cell* **74**, 1089–1100 (1993).
99. Ohashi, P. S. *et al.* Ablation of "tolerance" and induction of diabetes by virus infection in viral antigen transgenic mice. *Cell* **65**, 305–317 (1991).
100. Sarvetnick, N., Liggitt, D., Pitts, S. L., Hansen, S. E. & Stewart, T. A. Insulin-dependent diabetes mellitus induced in transgenic mice by ectopic expression of class II MHC and interferon- γ . *Cell* **52**, 773–782 (1988).

Acknowledgements

We apologize to the many investigators whose work we could not cite due to a limit on the number of references. We thank C. R. Kahn for his insights into type 2 diabetes. Our work on β -cell death is funded by a research grant from the Juvenile Diabetes Research Foundation.

Insulin signalling and the regulation of glucose and lipid metabolism

Alan R. Saltiel* & C. Ronald Kahn†

*Life Sciences Institute, Department of Medicine, University of Michigan School of Medicine, Ann Arbor, Michigan 48109, USA (e-mail: saltiel@umich.edu)

†Joslin Diabetes Centre, Harvard Medical School, Boston, Massachusetts 02215, USA (e-mail: c.ronald.kahn@joslin.harvard.edu)

The epidemic of type 2 diabetes and impaired glucose tolerance is one of the main causes of morbidity and mortality worldwide. In both disorders, tissues such as muscle, fat and liver become less responsive or resistant to insulin. This state is also linked to other common health problems, such as obesity, polycystic ovarian disease, hyperlipidaemia, hypertension and atherosclerosis. The pathophysiology of insulin resistance involves a complex network of signalling pathways, activated by the insulin receptor, which regulates intermediary metabolism and its organization in cells. But recent studies have shown that numerous other hormones and signalling events attenuate insulin action, and are important in type 2 diabetes.

Despite periods of feeding and fasting, plasma glucose remains in a narrow range between 4 and 7 mM in normal individuals. This tight control is governed by the balance between glucose absorption from the intestine, production by the liver and uptake and metabolism by peripheral tissues. Insulin increases glucose uptake in muscle and fat (see Box 1), and inhibits hepatic glucose production, thus serving as the primary regulator of blood glucose concentration. Insulin also stimulates cell growth and differentiation, and promotes the storage of substrates in fat, liver and muscle by stimulating lipogenesis, glycogen and protein synthesis, and inhibiting lipolysis, glycogenolysis and protein breakdown (Fig. 1). Insulin resistance or deficiency results in profound dysregulation of these processes, and produces elevations in fasting and postprandial glucose and lipid levels.

Insulin increases glucose uptake in cells by stimulating the translocation of the glucose transporter GLUT4 from intracellular sites to the cell surface (see Box 1). Up to 75% of insulin-dependent glucose disposal occurs in skeletal muscle, whereas adipose tissue accounts for only a small fraction¹. Despite this, mice with a knockout of the insulin receptor in muscle have normal glucose tolerance², whereas those with a knockout of the insulin-sensitive glucose transporter in fat have impaired glucose tolerance, apparently owing to insulin resistance being induced in muscle and liver³. Both obesity and lipoatrophy also cause insulin resistance and predisposition to type 2 diabetes, demonstrating that adipose tissue is crucial in regulating metabolism beyond its ability to take up glucose⁴. Although insulin does not stimulate glucose uptake in liver, it blocks glycogenolysis and gluconeogenesis, and stimulates glycogen synthesis, thus regulating fasting glucose levels. Insulin action in tissues not normally considered insulin sensitive, including brain and pancreatic β -cell, may also be important in glucose homeostasis^{2,5} (see below).

Proximal insulin-signalling pathways

The insulin receptor

The insulin receptor belongs to a subfamily of receptor tyrosine kinases that includes the insulin-like growth factor (IGF)-I receptor and the insulin receptor-related receptor (IRR)⁶. These receptors are tetrameric proteins consisting of

two α - and two β -subunits that function as allosteric enzymes in which the α -subunit inhibits the tyrosine kinase activity of the β -subunit. Insulin binding to the α -subunit leads to derepression of the kinase activity in the β -subunit followed by transphosphorylation of the β -subunits and a conformational change that further increases kinase activity⁶. Insulin, IGF-I and the IRR receptors can form functional hybrids; thus, an inhibitory mutation in one receptor can inhibit the activity of the others⁷.

Homologues of the insulin/IGF-I receptor have been identified in *Drosophila*, *Caenorhabditis elegans* and metazoan marine sponges⁸. These lower organisms use some of the same downstream signals critical to the regulation of mammalian cells, including phosphatidylinositol-3-OH kinase (PI(3)K), Akt and forkhead transcription factors. Inhibitory mutants of the insulin/IGF system in *C. elegans* live longer than normal animals, raising a number of interesting questions about the association of hyperinsulinaemia/insulin resistance with conditions that shorten life span, such as obesity, diabetes and accelerated atherosclerosis.

Insulin-receptor substrates

At least nine intracellular substrates of the insulin/IGF-I receptor kinases have been identified (Fig. 2). Four of these belong to the family of insulin-receptor substrate (IRS) proteins⁹. Other substrates include Gab-1, p60^{dok}, Cbl, APS and isoforms of Shc¹⁰. The phosphorylated tyrosines in these substrates act as 'docking sites' for proteins that contain SH2 (Src-homology-2) domains. Many of these SH2 proteins are adaptor molecules, such as the p85 regulatory subunit of PI(3)K and Grb2, or CrkII, which activate small G proteins by binding to nucleotide exchange factors. Others are themselves enzymes, including the phosphotyrosine phosphatase SHP2 and the cytoplasmic tyrosine kinase Fyn. Substrate binding to these SH2 proteins can regulate their activities, or in some cases their subcellular location.

Although the IRS proteins are highly homologous, recent studies in knockout mice and cell lines suggest that they serve complementary, rather than redundant, roles in insulin/IGF-I signalling. IRS-1-knockout mice exhibit generalized pre- and post-natal growth retardation, as well as insulin resistance in peripheral tissues and impaired glucose tolerance^{11,12}. IRS-2-knockout mice also exhibit insulin

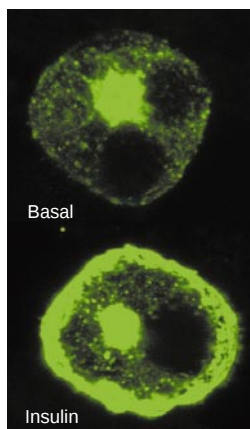
Box 1

The regulation of glucose transport

Insulin increases glucose transport in fat and muscle cells by stimulating the translocation of the transporter GLUT4 from intracellular sites to the plasma membrane (see figure below). GLUT4 is found in vesicles that continuously cycle from intracellular stores to the plasma membrane. Insulin increases glucose transport by increasing the rate of GLUT4-vesicle exocytosis, and by slightly decreasing the rate of internalization⁹⁹. Although the exact mechanisms are unknown, it is likely that the insulin-responsive GLUT4 vesicle is tethered to intracellular sites, perhaps defined by a microtubule network¹⁰⁰. Phosphorylation events probably catalysed by PtdIns(3,4,5)P₃-dependent kinases release vesicles from these sites, allowing trafficking of GLUT4 to the cell surface. Recent evidence suggests that these vesicles move along microtubule tracks en route to the cell surface, perhaps via kinesin motors. The vesicles then dock and fuse with the plasma membrane, allowing the extracellular exposure of the GLUT4 protein.

It is likely that the actin cytoskeleton is also crucial in insulin-stimulated GLUT4 translocation. Insulin causes remodelling of cortical actin filaments just below the plasma membrane, and induces membrane ruffling. This effect probably reflects actin polymerization and depolymerization involving lamellipodia and/or filopodia formation. The actin depolymerizing agent cytochalasin D, as well as the actin monomer-binding toxins latrunculin A and B, partially inhibit the effects of insulin on actin and GLUT4 translocation.

The docking and fusion of the GLUT4 vesicle at the plasma membrane may also be subject to regulation by insulin. The v-SNARE protein VAMP2 physically interacts with its t-SNARE counterpart syntaxin 4 on GLUT4 vesicles during vesicle docking and fusion with the plasma membrane⁹⁹. Although these SNARE interactions are essential, neither SNARE protein seems to be a direct target of insulin. Instead, the SNARE accessory proteins Synip and Munc18c may be involved in the control of GLUT4 docking and fusion in an insulin-dependent, PI(3)K-independent manner.



Box 1 Figure 3T3L1 adipocytes were transfected with a fusion construct of GLUT4 and enhanced green fluorescent protein. Cells were treated with or without insulin. The image is a confocal micrograph of single cells, showing the translocation of the GLUT4 protein from intracellular sites to the cell surface. For a three-dimensional image of these cells, visit <http://www.medgen.med.umich.edu/labs/saltiel/>.

resistance in both peripheral tissues and liver, but have defective growth in only some tissues, including certain regions of the brain, islets and retina^{13,14}. In the IRS-2^{-/-} mouse, this multifactorial insulin resistance combined with decreased β -cell mass leads to development of type 2 diabetes¹⁴. By contrast, IRS-3- and IRS-4-knockout mice have normal or near normal growth and metabolism¹⁵.

The different IRS proteins seem to serve different functions at the cellular level, probably owing to differences in tissue distribution, subcellular localization and intrinsic activity of the proteins. IRS-1-knockout cells exhibit reduced IGF-I-stimulated DNA synthesis^{11,12} and fail to differentiate into adipocytes in culture¹⁶. Likewise, the mitogenic response mediated by IRS-2 is weaker than that by produced by IRS-1 (ref. 17). IRS-2-knockout cells show a

major defect in insulin-stimulated glucose transport¹⁴. The roles of IRS-3 and 4 are less clear in cultured cells, but some data suggest that these substrates may act as negative regulators of IRS-1 and -2 (ref. 18).

Inhibition of insulin-receptor signalling

In addition to tyrosine phosphorylation, both the insulin receptor and IRS proteins undergo serine phosphorylation, which may attenuate signalling by decreasing insulin-stimulated tyrosine phosphorylation¹⁹ and promote interaction with 14-3-3 proteins²⁰. These inhibitory phosphorylations provide negative feedback to insulin signalling and serve as a mechanism for cross-talk from other pathways that produce insulin resistance. Several kinases have been implicated in this process, including PI(3)K, Akt, glycogen synthase kinase (GSK)-3 and mammalian target of rapamycin (mTOR). Recent data indicate that obesity-induced attenuation of insulin signalling might arise from the sequential activation of protein kinase C (PKC) and inhibitor of nuclear factor- κ B (I κ B) kinase, although the details of this pathway have not been elucidated^{21,22}.

Insulin action is also attenuated by protein tyrosine phosphatases (PTPases), which catalyse the rapid dephosphorylation of the receptor and its substrates. A number of PTPases have been identified that catalyse dephosphorylation of the insulin receptor *in vitro*, some of which are expressed in insulin-responsive cells, or upregulated in states of insulin resistance. Most attention has focused on the cytoplasmic phosphatase PTP1B. Knockout of PTP1B leads to increased tyrosine phosphorylation of the insulin receptor and IRS proteins in muscle and improved insulin sensitivity²³. PTP1B^{-/-} mice are also resistant to diet-induced obesity, suggesting the brain as an important site of action. This combination of effects implicates PTP1B as a potential therapeutic target in diabetes and obesity.

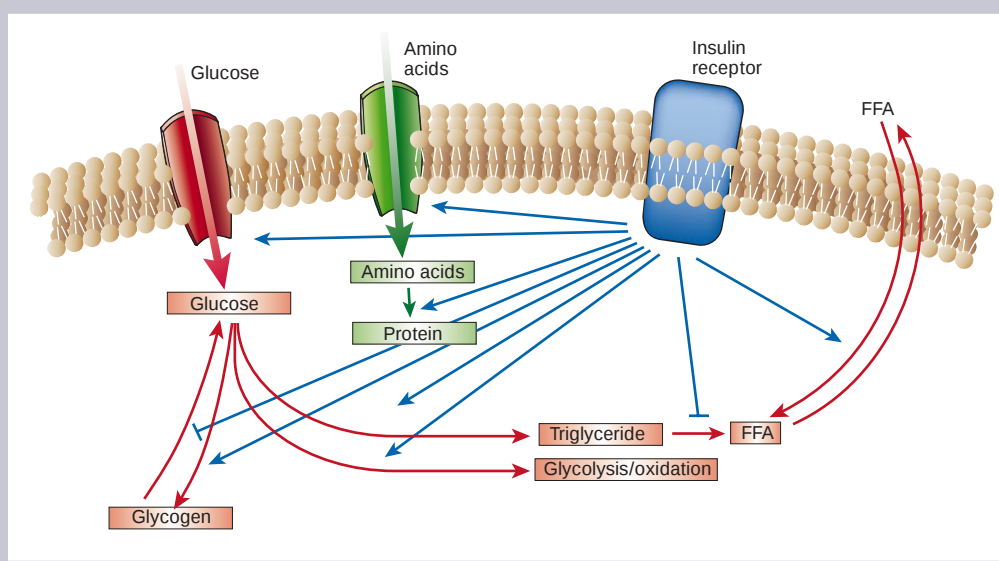
PI(3)K and insulin action

PI(3)K has a pivotal role in the metabolic and mitogenic actions of insulin and IGF-I (ref. 24). Inhibitors of class Ia PI(3)K, or transfections with dominant negative constructs of the enzyme, block most metabolic actions of insulin, including stimulation of glucose transport, glycogen and lipid synthesis. PI(3)K consists of a p110 catalytic subunit and a p85 regulatory subunit that possesses two SH2 domains that interact with tyrosine-phosphorylated pYMXM and pYXXM motifs in IRS proteins²⁵. At least eight isoforms of the regulatory subunits have been identified. These are derived from three genes (*p85 α* , *p85 β* and *P55^{PIK}*)²⁶ and alternative splicing of *p85 α* to produce *AS53/p55 α* ²⁷ and *p50 α* ²⁸. Of these, *p85 α* is predominant and thought to be the main response pathway for most stimuli.

The exact roles of the different regulatory subunits of PI(3)K in insulin action are unclear. Splice variants possess differences in potencies for enzyme activation²⁷, tissue distribution^{27,28} and sensitivity to insulin²⁹. Knockout mice with a disruption of all three isoforms derived from the *p85 α* gene die shortly after birth³⁰, whereas heterozygous knockout mice or mice lacking only full-length *p85 α* are viable and exhibit improved insulin sensitivity (ref. 31, and F. Mauvais-Jarvis *et al.*, personal communication). Cell lines derived from heterozygous knockouts also exhibit increased insulin/IGF-I signalling, which seem to be due to improved stoichiometry of the interaction (see below).

The activation of PI(3)K may transmit multiple signals. PI(3)K catalyses the phosphorylation of phosphoinositides on the 3-position to produce phosphatidylinositol-3-phosphates, especially PtdIns(3,4,5)P₃, which bind to the pleckstrin homology (PH) domains of a variety of signalling molecules thereby altering their activity or subcellular localization³². Moreover, PI(3)K also possesses serine kinase activity, and both the regulatory and catalytic subunits of the enzyme can interact with other signalling proteins. Indeed, recent studies suggest that these proteins may be important in insulin action independent of PtdIns(3,4,5)P₃ generation³³.

Figure 1 The regulation of metabolism by insulin. Insulin is the most potent anabolic hormone known, and promotes the synthesis and storage of carbohydrates, lipids and proteins, while inhibiting their degradation and release into the circulation. Insulin stimulates the uptake of glucose, amino acids and fatty acids into cells, and increases the expression or activity of enzymes that catalyse glycogen, lipid and protein synthesis, while inhibiting the activity or expression of those that catalyse degradation.



Phosphatidylinositol-3-phosphates regulate three main classes of signalling molecules: the AGC family of serine/threonine protein kinases³⁴, guanine nucleotide-exchange proteins of the Rho family of GTPases³⁵, and the TEC family of tyrosine kinases³⁶. PI(3)K also activates the mTOR/FRAP pathway, and might be involved in regulation of phospholipase D, leading to hydrolysis of phosphatidylcholine and increases in phosphatidic acid and diacylglycerol. The best characterized of the AGC kinases is phosphoinositide-dependent kinase 1 (PDK1), one of the serine kinases that phosphorylates and activates the serine/threonine kinase Akt/PKB³⁷. Akt possesses a PH domain that also interacts directly with PtdIns(3,4,5)P₃, promoting membrane targeting of the protein and catalytic activation. Akt has been suggested to be important in transmission of the insulin signal, by phosphorylation of the enzyme GSK-3 (see below), the forkhead transcription factors and cAMP response element-binding protein^{38,39}. Although studies using inhibitory or activated forms of Akt have not uniformly inhibited or mimicked insulin actions⁴⁰, deletion of Akt2 produces hepatic insulin resistance in mice⁴¹. Other

AGC kinases that are downstream of PI(3)K include serum- and glucocorticoid-regulated kinase and the atypical PKCs, PKC- ζ and - λ ⁴². Akt and/or the atypical PKCs seem to be required for insulin-stimulated glucose transport.

Activity of this pathway is also determined by phosphatidylinositol-3-phosphates such as phosphatase and tensin homologue⁴³ and the SH2 domain-containing inositol-5-phosphatase SHIP2 (ref. 44). Overexpression of these enzymes leads to decreased levels of PtdIns(3,4,5)P₃. This might terminate signal transduction and/or change the nature of the phosphoinositides, altering the binding specificity to PH or phox homology domains. Disruption of these genes or reducing expression of these messenger RNAs yields mice with increased insulin sensitivity⁴⁵.

The CAP/Cbl pathway and lipid rafts

In addition to PI(3)K activity, other signals seem to be required for insulin-stimulated glucose uptake¹⁰. This second pathway appears to involve tyrosine phosphorylation of the Cbl protooncogene⁴⁶. In

Figure 2 Signal transduction in insulin action. The insulin receptor is a tyrosine kinase that undergoes autophosphorylation, and catalyses the phosphorylation of cellular proteins such as members of the IRS family, Shc and Cbl. Upon tyrosine phosphorylation, these proteins interact with signalling molecules through their SH2 domains, resulting in a diverse series of signalling pathways, including activation of PI(3)K and downstream PtdIns(3,4,5)P₃-dependent protein kinases, ras and the MAP kinase cascade, and Cbl/CAP and the activation of TC10. These pathways act in a concerted fashion to coordinate the regulation of vesicle trafficking, protein synthesis, enzyme activation and inactivation, and gene expression, which results in the regulation of glucose, lipid and protein metabolism.

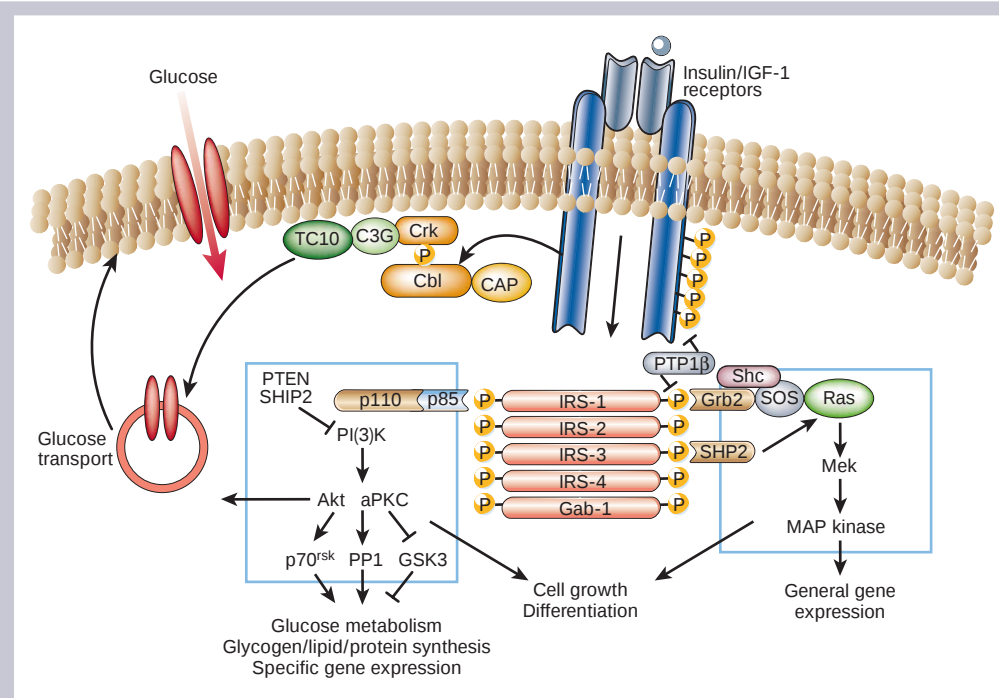
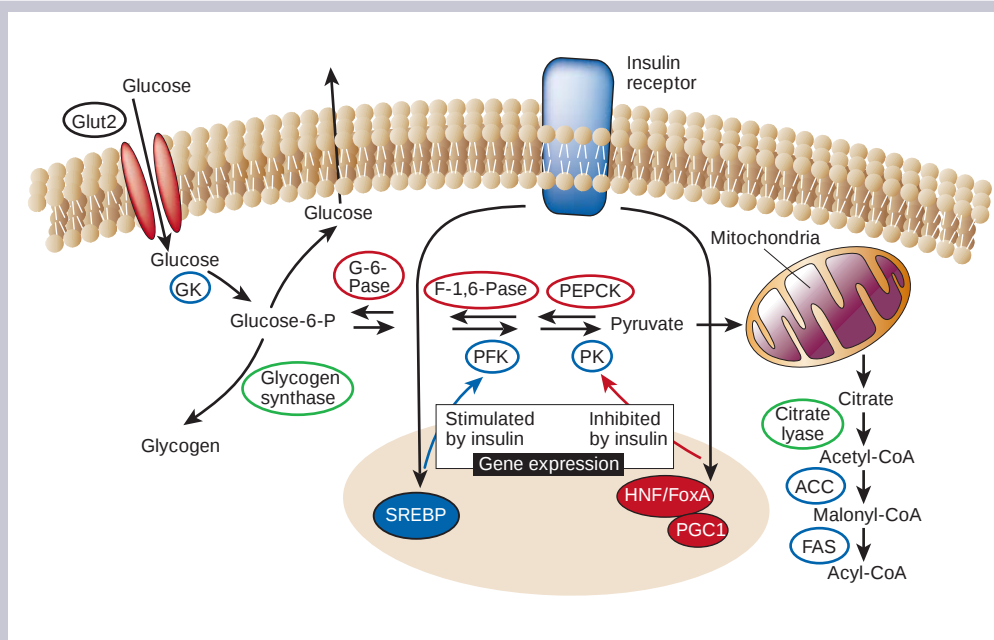


Figure 3 The regulation of glucose metabolism in the liver. In the hepatocyte, insulin stimulates the utilization and storage of glucose as lipid and glycogen, while repressing glucose synthesis and release. This is accomplished through a coordinated regulation of enzyme synthesis and activity. Insulin stimulates the expression of genes encoding glycolytic and fatty-acid synthetic enzymes (in blue), while inhibiting the expression of those encoding gluconeogenic enzymes (in red). These effects are mediated by a series of transcription factors and co-factors, including sterol regulatory element-binding protein (SREBP)-1, hepatic nuclear factor (HNF)-4, the forkhead protein family (Fox) and PPAR- γ co-activator 1 (PGC1). The hormone also regulates the activities of some enzymes, such as glycogen synthase and citrate lyase (in green), through changes in phosphorylation state. GK, glucokinase; Glucose-6-P, glucose-6-phosphate; G-6-Pase, glucose-6-phosphatase; F-1,6-Pase, fructose-1,6-bisphosphatase;



PEPCK, phosphoenolpyruvate carboxykinase; PFK, phosphofructokinase; PK, pyruvate kinase; ACC, acetyl-CoA carboxylase; FAS, fatty-acid synthase.

most insulin-responsive cells, Cbl is associated with the adapter protein CAP, which binds to proline-rich sequences in Cbl through its carboxyl-terminal SH3 domain⁴⁷. CAP is expressed in insulin-sensitive tissues, is markedly induced during adipocyte differentiation and its expression is increased by insulin-sensitizing peroxisome proliferator-activated receptor- γ (PPAR γ) agonists⁴⁸.

CAP belongs to a family of adapter proteins with a common organization containing three SH3 domains and a region of similarity to the peptide sorbin, referred to as a sorbin homology (SoHo) domain⁴⁹. Upon phosphorylation, the Cbl-CAP complex translocates to lipid raft domains in the plasma membrane, mediated by the interaction of the SoHo domain of CAP with the protein flotillin⁵⁰. Expression of dominant-interfering CAP mutants that cannot bind to Cbl or flotillin inhibit Cbl translocation and insulin-stimulated glucose uptake. Translocation of phosphorylated Cbl recruits the adapter protein CrkII to the lipid raft via interaction of the SH2 domain of CrkII with phospho-Cbl⁵¹. CrkII also forms a constitutive complex with the guanyl nucleotide-exchange protein C3G. Once translocated into lipid rafts, C3G comes into proximity with the G protein TC10, and catalyses the exchange of GTP for GDP, resulting in the activation of the protein⁵¹. The localization of TC10 in rafts is required for its activation by insulin⁵². Once activated, TC10 seems to provide a second signal to the GLUT4 protein that functions in parallel with the activation of the PI(3)K pathway⁵¹. This may involve the stabilization of cortical actin, which seems to be important in GLUT4 vesicle translocation to the plasma membrane (see Box 1).

Insulin-stimulated phosphorylation cascades

As is the case for other growth factors, insulin stimulates the mitogen-activated protein (MAP) kinase extracellular signal-regulated kinase (ERK) (Fig. 2). This pathway involves the tyrosine phosphorylation of IRS proteins and/or Shc, which in turn interact with the adapter protein Grb2, recruiting the Son-of-sevenless (SOS) exchange protein to the plasma membrane for activation of Ras. The activation of Ras also requires stimulation of the tyrosine phosphatase SHP2, through its interaction with receptor substrates such as Gab-1 or IRS1/2. Once activated, Ras operates as a molecular switch, stimulating a serine kinase cascade through

the stepwise activation of Raf, MEK and ERK. Activated ERK can translocate into the nucleus, where it catalyses the phosphorylation of transcription factors such as p62^{TCF}, initiating a transcriptional programme that leads to cellular proliferation or differentiation⁵³. Blockade of the pathway with dominant negative mutants or pharmacological inhibitors prevents the stimulation of cell growth by insulin, but has no effect on the metabolic actions of the hormone⁵⁴.

Insulin increases synthesis and blocks the degradation of proteins through activation of mTOR. mTOR is a member of the PI(3)K family of proteins, but seems to act primarily as a serine, rather than lipid kinase⁵⁵. The stimulation of this protein kinase involves PI(3)K activation, although another signal may also be required⁵⁶. mTOR can control the mammalian translation machinery by direct phosphorylation and activation of p70 ribosomal S6 kinase (p70^{rsk})⁵⁷, as well as phosphorylation of the initiation factor 4E for eukaryotic translation (eIF-4E) inhibitor, PHAS1 or 4E-binding protein 1 (ref. 58). P70^{rsk} activates ribosome biosynthesis by phosphorylating the ribosomal S6 protein, producing increased translation of mRNAs with a 5'-terminal oligopyrimidine tract. P70^{rsk} also requires a second PtdIns(3,4,5)P₃-dependent phosphorylation, presumably catalysed by PDK1. Phosphorylation of PHAS-1 by mTOR results in its dissociation from eIF-2, allowing cap-dependent translation of mRNAs with a highly structured 5'-untranslated region. Although the mechanism of activation of mTOR remains unclear, it seems to require the presence of amino acids in the media for full activation by growth factors, and thus may also represent a nutrient sensor⁵⁵.

Glucose and lipid regulation

Regulation of glycogen synthesis

Insulin stimulates glycogen accumulation through a coordinated increase in glucose transport and glycogen synthesis. The hormone activates glycogen synthase by promoting its dephosphorylation, through the inhibition of kinases such as PKA or GSK-3 (ref. 59), and activation of protein phosphatase 1 (PP1)⁶⁰. Upon its activation downstream of PI(3)K, Akt phosphorylates and inactivates GSK-3, decreasing the rate of phosphorylation of glycogen synthase, thus increasing its activity state⁵⁹. Insulin does not activate

PP1 globally, but rather specifically targets discrete pools of the phosphatase, primarily increasing PP1 activity localized at the glycogen particle. The compartmentalized activation of PP1 by insulin is due to glycogen-targeting subunits, which serve as 'molecular scaffolds', bringing together the enzyme directly with its substrates glycogen synthase and phosphorylase in a macromolecular complex, and in the process exerting profound effects on PP1 substrate-specific activity⁶¹.

Four different proteins have been reported to target PP1 to the glycogen particle. Despite a proposed common function, no two targeting subunits share more than 50% sequence homology, and this is largely confined to the PP1- and glycogen-binding regions. Overexpression of these scaffolding proteins in cells or *in vivo* results in a marked increase in cellular glycogen levels⁶¹. Although the mechanism by which insulin activates glycogen-associated PP1 remains unknown, inhibitors of PI(3)K block this effect, suggesting that PtdIns(3,4,5)P₃-dependent protein kinases are involved. These scaffolding proteins have a critical permissive role in the hormonal activation of the enzyme, perhaps interacting with additional proteins that regulate the interaction of PP1 with glycogen synthase and phosphorylase.

Regulation of gluconeogenesis

Insulin inhibits the production and release of glucose by the liver by blocking gluconeogenesis and glycogenolysis (Fig. 3). This occurs through a direct effect of insulin on the liver⁶², as well as by indirect effects of insulin on substrate availability⁶³. Insulin can also influence glucose metabolism indirectly by changes in free fatty acids generated from visceral fat, the so called 'single gateway' hypothesis⁶⁴. Because visceral fat is less sensitive to insulin than subcutaneous fat, even after a meal there is little suppression of lipolysis by the hormone in this fat depot. The resulting direct flux of fatty acids derived from these fat cells through the portal vein to the liver can stimulate glucose production, thus providing a signal for both insulin action and insulin resistance in the liver.

Insulin directly controls the activities of a set of metabolic enzymes by phosphorylation or dephosphorylation and also regulates the expression of genes encoding hepatic enzymes of gluconeogenesis and glycolysis⁶⁵. It inhibits the transcription of the gene encoding phosphoenolpyruvate carboxylase, the rate-limiting step in gluconeogenesis⁶⁶. The hormone also decreases transcription of the genes encoding fructose-1,6-bisphosphatase and glucose-6-phosphatase, and increases transcription of glycolytic enzymes such as glucokinase and pyruvate kinase, and lipogenic enzymes such as fatty acid synthase and acetyl-CoA carboxylase. Although the transcription factors that control the expression of these genes have remained elusive, new data suggest a potential role for the

forkhead family of transcription factors through phosphorylation by Akt-related protein kinases³⁹, and the PPAR γ co-activator PGC-1 (ref. 67).

Regulation of lipid synthesis and degradation

As is the case with carbohydrate metabolism, insulin also promotes the synthesis of lipids, and inhibits their degradation. Recent studies suggest that many of these changes require an increase in the transcription factor steroid regulatory element-binding protein (SREBP)-1c⁶⁸. Dominant negative forms of SREBP-1 can block expression of these gluconeogenic and lipogenic genes⁶⁹, whereas overexpression can increase their transcription⁶⁸. Hepatic SREBP levels are increased in some rodent models of lipodystrophy, and this is coordinated with increases in fatty acid synthesis and gluconeogenesis, the exact phenotype observed in genetic models of obesity-induced diabetes⁷⁰. Thus, increased expression of SREBP-1c might contribute to the insulin resistance observed in liver of diabetic rodents, with increased rates of both gluconeogenesis and lipogenesis. The pathways that account for the changes in SREBP-1c expression in response to insulin or other metabolic changes are not known, but probably lie downstream of the IRS/PI(3)K pathway.

In adipocytes, glucose is stored primarily as lipid, owing to increased uptake of glucose and activation of lipid synthetic enzymes, including pyruvate dehydrogenase, fatty acid synthase and acetyl-CoA carboxylase. Insulin also profoundly inhibits lipolysis in adipocytes, primarily through inhibition of the enzyme hormone-sensitive lipase⁷¹. This enzyme is acutely regulated by control of its phosphorylation state, which is activated by PKA-dependent phosphorylation, and inhibited as a result of a combination of kinase inhibition and phosphatase activation. Insulin inhibits the activity of the lipase primarily through reductions in cAMP levels, owing to the activation of a cAMP-specific phosphodiesterase in fat cells⁷².

What causes insulin resistance?

The insulin resistance of obesity and type 2 diabetes is characterized by defects at many levels, with decreases in receptor concentration and kinase activity, the concentration and phosphorylation of IRS-1 and -2, PI(3)K activity, glucose transporter translocation, and the activity of intracellular enzymes¹⁰. Activation of the MAP kinase pathway by insulin is not reduced in type 2 diabetes, perhaps allowing for some of the detrimental effects of chronic hyperinsulinaemia on cellular growth in the vasculature⁷³.

Genetic and acquired factors can profoundly influence insulin sensitivity. Genetic defects in the insulin receptor are relatively rare, but represent the most severe forms of insulin resistance, and are exemplified by leprechaunism, the Rabson Mendenhall Syndrome, and the type A syndrome of insulin resistance⁷⁴. Differences in clinical presentation may be due to the severity of the genetic defect, the ability of the mutant receptors to form hybrids with IGF-I or other receptors, and other background genetic or acquired factors that modify the insulin-resistant state. Type 2 diabetes is polygenic and may involve polymorphisms in multiple genes encoding the proteins involved in insulin signalling, insulin secretion and intermediary metabolism⁷⁵.

Targeted deletions of the components of insulin signalling *in vivo* using homologous recombination have yielded some insight into the complexity of these mechanisms. Although some single defects in the insulin-signalling pathway, such as knockout of the insulin receptor IRS-2 or Akt2, can produce diabetes, knockout of the p85 subunit of PI(3)K, IRS-1 or GLUT4 does not (summarized in Table 1). Conversely, knockout of single genes that are involved in turning off the insulin signal, such as PTP1B and SHIP2, ameliorate diabetes in obese rodents^{23,45}.

Combinatorial knockouts have been produced that mimic polygenic type 2 diabetes with heterozygous deletion of the insulin receptor and IRS-1 (ref. 76), of the insulin receptor, IRS-1 and IRS-2 (ref. 13), and of IRS-1 and glucokinase⁷⁷. In some of these combina-

Table 1 Mice phenotypes with single-gene knockouts in signalling pathways

Gene	Phenotype	Reference
Insulin receptor	Normal intrauterine growth; die of diabetic ketoacidosis at 3–7 days	76
IGF1 receptor	Intrauterine and postnatal growth retardation; normal glucose homeostasis	97
IRS-1	Insulin resistance/impaired glucose tolerance; IGF resistance/growth retardation	11,12
IRS-2	Insulin resistance/decreased β -cell development; type 2 diabetes	13,14
IRS-3	Normal growth/normal glucose tolerance	15
IRS-4	Normal growth/normal glucose tolerance	15
Akt2	Insulin resistance in liver and muscle	41
GLUT4	Cardiac hypertrophy/failure; normal glucose tolerance	81
P85 α (hetero)	Increased insulin sensitivity; hypoglycaemia	30,31
PTP1B	Increased insulin sensitivity; resistance to diet-induced obesity	23
SHIP2	Increased insulin sensitivity	45

Table 2 Defects in tissue-specific knockout mice

Gene	Tissue	Phenotype	Reference
Insulin receptor	Skeletal muscle	Normal glucose tolerance; increased fat mass; increased triglycerides and FFAs	2,82
	Liver	Impaired glucose tolerance; hyperinsulinaemia and reduced insulin clearance; decreased hepatic function	62
	β -cell	Loss of glucose-stimulated insulin secretion; progressively impaired glucose tolerance; decreased β -cell growth in adults	5
	Brain	Increased appetite; increased fat and leptin; insulin resistance; hypothalamic hypogonadism	79
GLUT4 glucose transporter	Skeletal muscle	Reduced basal, insulin and contraction-stimulated glucose transport; severe insulin resistance; glucose intolerance	80
	Fat	Impaired glucose tolerance; hyperinsulinaemia; secondary insulin resistance in muscle and liver	3
Glucokinase	β -cell	Die within a few days of birth with severe diabetes	98
	Liver	Mild hyperglycaemia; pronounced defects in glycogen synthesis and glucose turnover; impaired glucose-stimulated insulin secretion	98

tions there has been clear evidence of genetic epistasis. For example, although heterozygous knockout of either the insulin receptor or IRS-1 alone does not produce diabetes, the double-heterozygous knockout produces diabetes in up to 50% of mice. This striking finding gives insight into human type 2 diabetes, where insulin-induced downregulation or genetic polymorphisms in the receptor or IRS-1 alone might produce only modest changes in signalling capacity, but when combined can lead to diabetes.

One genetic model that produced a surprising phenotype regarding glucose homeostasis emerged from the knockout of the p85 α regulatory subunits of PI(3)K. Although, PI(3)K is central to the metabolic actions of insulin, p85 α heterozygous knockout mice counter-intuitively exhibit improved insulin sensitivity^{30,31}. Further-

more, superimposition of p85 α heterozygosity on the insulin receptor/IRS-1 double-heterozygous knockout protects against diabetes (Mauvais *et al.*, personal communication). This surprising protection seems to be due to a unique feature of the insulin-signalling pathway in which the stoichiometric balance between p85 α , the catalytic subunit p110 and IRS proteins is critical for optimal signal transduction.

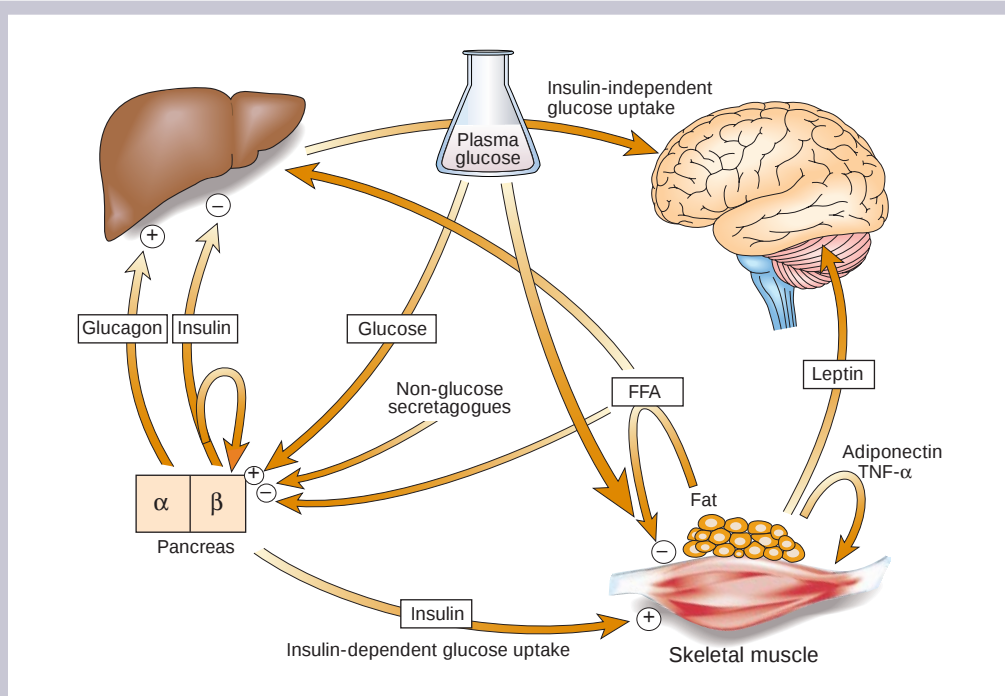
The role of specific tissues in the pathogenesis of diabetes has been explored using the *Cre-lox* DNA-recombination technology to create tissue-specific knockouts of the insulin receptor^{2,62,78,79} and GLUT4 (refs 3,80; see Table 2). Despite the absence of diabetes in mice with a global knockout of GLUT4 (ref. 81), tissue-specific knockouts of GLUT4 in muscle⁸⁰ and fat³ have resulted in severely impaired glucose tolerance. Tissue-specific knockout of the insulin receptor has also produced often surprising results. As noted above, despite recognition that insulin stimulates glucose uptake primarily in muscle, mice with a knockout of the muscle insulin receptor have normal glucose tolerance². This occurs, at least in part, as the result of a shift of glucose uptake into fat, with subsequent increases in adipose tissue mass, circulating free fatty acids (FFAs) and triglycerides⁸². Mice with a knockout of the fat-specific insulin receptor also have normal glucose tolerance, whereas the liver-specific insulin-receptor knockout shows both impaired glucose tolerance and decreased insulin clearance with marked hyperinsulinaemia⁶². Perhaps the most surprising results, however, have come from studies of mice with knockouts of the β -cell-specific insulin receptor and neural/brain-specific insulin receptor⁷⁹. The former exhibit a marked defect in glucose-stimulated insulin secretion similar to that observed in type 2 diabetes, whereas the latter exhibit increased food intake, mild adiposity, insulin resistance and hypertriglyceridaemia, as well as reduced fertility due to hypothalamic hypogonadism. Taken together, these findings suggest a unifying hypothesis for type 2 diabetes in which insulin resistance in classic target tissues, such as liver, muscle and fat, coupled with insulin resistance in β -cell, brain and other tissues, combine to produce the pathophysiology of type 2 diabetes.

The fat cell regulates insulin sensitivity

Free fatty acids

Adipose tissue has a special role in insulin resistance (Fig. 4). Circulating FFAs derived from adipocytes are elevated in many insulin-resistant states and have been suggested to contribute to the

Figure 4 Cross-talk between tissues in the regulation of glucose metabolism. Insulin is secreted from the β -cells of the pancreas in response to elevations in plasma glucose. The hormone decreases glucose production from the liver, and increases glucose uptake, utilization and storage in fat and muscle. The fat cell is important in metabolic regulation, releasing FFAs that reduce glucose uptake in muscle, insulin secretion from the β -cell, and increase glucose production from the liver. The fat cell can also secrete 'adipokines' such as leptin, adiponectin and TNF- α , which regulate food intake, energy expenditure and insulin sensitivity.



insulin resistance of diabetes and obesity by inhibiting glucose uptake, glycogen synthesis and glucose oxidation, and by increasing hepatic glucose output⁶³. Elevated FFAs are also associated with a reduction in insulin-stimulated IRS-1 phosphorylation and IRS-1-associated PI(3)K activity⁸³. The link between increased circulating FFAs and insulin resistance might involve accumulation of triglycerides and fatty acid-derived metabolites (diacylglycerol, fatty acyl-CoA and ceramides) in muscle and liver. Nuclear magnetic resonance spectroscopy has shown a close correlation between intramyocellular triglyceride content and whole-body insulin resistance in patients with obesity and type 2 diabetes⁸⁴. Transgenic mice with muscle- or liver-specific overexpression of lipoprotein lipase exhibit increases in tissue triglyceride content that correlate with decreases in insulin action and activation of IRS-associated PI(3)K⁸⁵. Activation of PKC and/or I κ B kinase and serine phosphorylation of the insulin receptor and its substrates might be important⁸³.

Adipokines

In addition to its role as a storage depot for lipid, the fat cell produces and secretes a number of hormones, collectively called adipokines, which may profoundly influence metabolism and energy expenditure. Expression of tumour-necrosis factor- α (TNF- α) is increased in fat of obese rodents and humans, and has been shown to produce serine phosphorylation of IRS-1, resulting in reduced insulin receptor kinase activity and insulin resistance¹⁹. In rodents, anti-TNF- α reagents significantly improved insulin resistance⁸⁶, although in humans the importance of this mechanism is much debated, as limited studies of anti-TNF reagents have shown little or no effect on the insulin-resistant state⁸⁷.

Leptin is a member of the cytokine family of hormones that is produced by adipose tissue and acts on receptors in the central nervous system and other sites to inhibit food intake and promote energy expenditure. Insulin resistance characterizes states of severe leptin deficiency or resistance, such as *ob/ob* or *db/db* mice, or genetic models of lipotrophic diabetes. In some of these, administration of exogenous leptin improves glucose tolerance and insulin sensitivity independently of effects on food intake, probably by affecting neuroendocrine pathways that modulate insulin action in the liver^{88,89}, although this cytokine might also have direct effects on hepatic cells⁹⁰.

Adiponectin (also called Acrp30 or adipoQ) is a fat cell-derived peptide possessing a collagenous domain at the amino terminus and a globular domain that shares homology with complement factor C1q⁹¹. Recent studies have shown that expression of adiponectin mRNA is decreased in obese humans and mice and some models of lipotrophic diabetes. Acute treatment of mice with this adipokine decreases insulin resistance, decreases plasma FFAs and the triglyceride content of muscle and liver, and increases expression of genes involved in fatty-acid oxidation and energy expenditure⁹². In lipotrophic mice, insulin resistance is reversed by the combination of physiological doses of adiponectin and leptin, but only partially by either adiponectin or leptin alone. In isolated hepatocytes, adiponectin increased the ability of insulin to suppress glucose production⁹³. A recent genome-wide scan in humans also mapped a susceptibility locus for type 2 diabetes and metabolic syndrome to chromosome 3q27 in a region near the adiponectin gene⁹⁴.

Resistin is the most recently discovered peptide hormone to be secreted by adipocytes. Resistin belongs to a family of related secreted proteins termed RELMs (resistin-like molecules) and FIZZ (found in inflammatory zone)⁹⁵. Initial studies suggested that resistin might cause insulin resistance, as levels were increased in obese mice and reduced by antidiabetic drugs of the thiazolidinedione class. Furthermore, administration of antiresistin antibody seemed to improve blood sugar and insulin action in mice with diet-induced obesity. But subsequent studies have not confirmed these initial findings⁹⁶. The

potential role of resistin is further complicated by uncertainty about existence of a human homologue of the hormone.

Perspectives

There has been considerable progress over the past few years in unravelling the mechanisms of insulin action, and the molecular defects that give rise to insulin resistance. Recent advances dissecting the signalling pathways, cellular architecture and spatial compartmentalization of proteins, coupled with the sophisticated genetic analysis of the system, have yielded quantum leaps in our insight into how proteins and tissues interact to control glucose and lipid metabolism. But many gaps remain in our understanding of these processes, and in the pathophysiology underlying insulin resistance. We need to define the missing steps in the insulin-signalling network, elucidate the mechanisms of cross-talk in the system, and determine the genetic susceptibility of insulin resistance and the interactions between genes and environment. Such studies will provide new insights into diabetes and insulin resistance, perhaps even allowing a more focused and individualized approach to therapy or prevention of these disorders. □

- Klip, A. & Paquet, M. R. Glucose transport and glucose transporters in muscle and their metabolic regulation. *Diabetes Care* **13**, 228–243 (1990).
- Bruning, J. C. *et al.* A muscle-specific insulin receptor knockout exhibits features of the metabolic syndrome of NIDDM without altering glucose tolerance. *Mol. Cell* **2**, 559–569 (1998).
- Abel, E. D. *et al.* Adipose-selective targeting of the GLUT4 gene impairs insulin action in muscle and liver. *Nature* **409**, 729–733 (2001).
- Gavrilova, O. *et al.* Surgical implantation of adipose tissue reverses diabetes in lipotrophic mice. *J. Clin. Invest.* **105**, 271–278 (2000).
- Kulkarni, R. N. *et al.* Tissue-specific knockout of the insulin receptor in pancreatic beta cells creates an insulin secretory defect similar to that in type 2 diabetes. *Cell* **96**, 329–339 (1999).
- Patti, M. E. & Kahn, C. R. The insulin receptor—a critical link in glucose homeostasis and insulin action. *J. Basic Clin. Physiol. Pharmacol.* **9**, 89–109 (1998).
- Butler, A. A. & LeRoith, D. Tissue-specific versus generalized gene targeting of the *igf1* and *igf1r* genes and their roles in insulin-like growth factor physiology. *Endocrinology* **142**, 1685–1688 (2001).
- Skorokhod, A. *et al.* Origin of insulin receptor-like tyrosine kinases in marine sponges. *Biol. Bull.* **197**, 198–206 (1999).
- White, M. F. The IRS-signalling system: a network of docking proteins that mediate insulin action. *Mol. Cell. Biochem.* **182**, 3–11 (1998).
- Pessin, J. E. & Saltiel, A. R. Signaling pathways in insulin action: molecular targets of insulin resistance. *J. Clin. Invest.* **106**, 165–169 (2000).
- Tamemoto, H. *et al.* Insulin resistance and growth retardation in mice lacking insulin receptor substrate-1. *Nature* **372**, 182–186 (1994).
- Araki, E. *et al.* Alternative pathway of insulin signaling in mice with targeted disruption of the IRS-1 gene. *Nature* **372**, 186–190 (1994).
- Kido, Y. *et al.* Tissue-specific insulin resistance in mice with mutations in the insulin receptor, IRS-1, and IRS-2. *J. Clin. Invest.* **105**, 199–205 (2000).
- Withers, D. J. *et al.* Disruption of IRS-2 causes type 2 diabetes in mice. *Nature* **391**, 900–904 (1998).
- Fantini, V. R., Wang, Q., Lienhard, G. E. & Keller, S. R. Mice lacking insulin receptor substrate 4 exhibit mild defects in growth, reproduction, and glucose homeostasis. *Am. J. Physiol. Endocrinol. Metab.* **278**, E127–E133 (2000).
- Fasshauer, M. *et al.* Essential role of insulin receptor substrate 1 in differentiation of brown adipocytes. *Mol. Cell. Biol.* **21**, 319–329 (2001).
- Mendez, R., Myers, M. G. Jr, White, M. F. & Rhoads, R. E. Stimulation of protein synthesis, eukaryotic translation initiation factor 4E phosphorylation, and PHAS-I phosphorylation by insulin requires insulin receptor substrate 1 and phosphatidylinositol 3-kinase. *Mol. Cell. Biol.* **16**, 2857–2864 (1996).
- Tsuruzoe, K., Emkey, R., Kriauciunas, K. M., Ueki, K. & Kahn, C. R. Insulin receptor substrate 3 (IRS-3) and IRS-4 impair IRS-1- and IRS-2-mediated signaling. *Mol. Cell. Biol.* **21**, 26–38 (2001).
- Hotamisligil, G. S. *et al.* IRS-1-mediated inhibition of insulin receptor tyrosine kinase activity in TNF- α - and obesity-induced insulin resistance. *Science* **271**, 665–668 (1996).
- Craparo, A., Freund, R. & Gustafson, T. A. 14-3-3 (ϵ) interacts with the insulin-like growth factor I receptor and insulin receptor substrate 1 in a phosphoserine-dependent manner. *J. Biol. Chem.* **272**, 11663–11669 (1997).
- Yuan, M. *et al.* Reversal of obesity- and diet-induced insulin resistance with salicylates or targeted disruption of *Ikk β* . *Science* **293**, 1673–1677 (2001).
- Kim, J. K. *et al.* Prevention of fat-induced insulin resistance by salicylate. *J. Clin. Invest.* **108**, 437–446 (2001).
- Elchebly, M. *et al.* Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. *Science* **283**, 1544–1548 (1999).
- Shepherd, P. R., Nave, B. T. & Siddle, K. Insulin stimulation of glycogen synthesis and glycogen synthase activity is blocked by wortmannin and rapamycin in 3T3-L1 adipocytes: evidence for the involvement of phosphoinositide 3-kinase and p70 ribosomal protein-S6 kinase. *Biochem. J.* **305**, 25–28 (1995).
- Myers, M. G. Jr *et al.* IRS-1 activates phosphatidylinositol 3'-kinase by associating with src homology 2 domains of p85. *Proc. Natl Acad. Sci. USA* **89**, 10350–10354 (1992).
- Pons, S. *et al.* The structure and function of p55PIK reveal a new regulatory subunit for phosphatidylinositol 3-kinase. *Mol. Cell. Biol.* **15**, 4453–4465 (1995).
- Antonetti, D. A., Algenstaedt, P. & Kahn, C. R. Insulin receptor substrate 1 binds two novel splice variants of the regulatory subunit of phosphatidylinositol 3-kinase in muscle and brain. *Mol. Cell. Biol.* **16**, 2195–2203 (1996).

28. Fruman, D. A., Cantley, L. C. & Carpenter, C. L. Structural organization and alternative splicing of the murine phosphoinositide 3-kinase p85 alpha gene. *Genomics* **37**, 113–121 (1996).
29. Kerouz, N. J., Horsch, D., Pons, S. & Kahn, C. R. Differential regulation of insulin receptor substrates-1 and -2 (IRS-1 and IRS-2) and phosphatidylinositol 3-kinase isoforms in liver and muscle of the obese diabetic (ob/ob) mouse. *J. Clin. Invest.* **100**, 3164–3172 (1997).
30. Fruman, D. A. *et al.* Hypoglycaemia, liver necrosis and perinatal death in mice lacking all isoforms of phosphoinositide 3-kinase p85 α . *Nature Genet.* **26**, 379–382 (2000).
31. Terauchi, Y. *et al.* Increased insulin sensitivity and hypoglycaemia in mice lacking the p85 α subunit of phosphoinositide 3-kinase. *Nature Genet.* **21**, 230–235 (1999).
32. Lietzke, S. E. *et al.* Structural basis of 3-phosphoinositide recognition by pleckstrin homology domains. *Mol. Cell* **6**, 385–394 (2000).
33. Kessler, A., Uphues, I., Ouwens, D. M., Till, M. & Eckel, J. Diversification of cardiac insulin signaling involves the p85 alpha/beta subunits of phosphatidylinositol 3-kinase. *Am. J. Physiol. Endocrinol. Metab.* **280**, E65–E74 (2001).
34. Peterson, R. T. & Schreiber, S. L. Kinase phosphorylation: keeping it all in the family. *Curr. Biol.* **9**, R521–R524 (1999).
35. Mackay, D. J. & Hall, A. Rho GTPases. *J. Biol. Chem.* **273**, 20685–20688 (1998).
36. Ziegler, S. F., Bird, T. A., Schneringer, J. A., Schooley, K. A. & Baum, P. R. Molecular cloning and characterization of a novel receptor protein tyrosine kinase from human placenta. *Oncogene* **8**, 663–670 (1993).
37. Alessi, D. R. *et al.* Characterization of a 3-phosphoinositide-dependent protein kinase which phosphorylates and activates protein kinase B α . *Curr. Biol.* **7**, 261–269 (1997).
38. Cross, D. A. *et al.* The inhibition of glycogen synthase kinase-3 by insulin or insulin-like growth factor 1 in the rat skeletal muscle cell line L6 is blocked by wortmannin, but not by rapamycin: evidence that wortmannin blocks activation of the mitogen-activated protein kinase pathway in L6 cells between Ras and Raf. *Biochem. J.* **303**, 21–26 (1994).
39. Nakae, J., Park, B. C. & Accili, D. Insulin stimulates phosphorylation of the forkhead transcription factor FKHR on serine 253 through a Wortmannin-sensitive pathway. *J. Biol. Chem.* **274**, 15982–15985 (1999).
40. Brady, M. J., Bourbonnais, F. J. & Saltiel, A. R. The activation of glycogen synthase by insulin switches from kinase inhibition to phosphatase activation during adipogenesis in 3T3-L1 cells. *J. Biol. Chem.* **273**, 14063–14066 (1998).
41. Cho, H. *et al.* Insulin resistance and a diabetes mellitus-like syndrome in mice lacking the protein kinase Akt2 (PKB β). *Science* **292**, 1728–1731 (2001).
42. Standaert, M. L. *et al.* Protein kinase C- ζ as a downstream effector of phosphatidylinositol 3-kinase during insulin stimulation in rat adipocytes. Potential role in glucose transport. *J. Biol. Chem.* **272**, 30075–30082 (1997).
43. Ogg, S. & Ruvkun, G. The C. elegans PTEN homolog, DAF-18, acts in the insulin receptor-like metabolic signaling pathway. *Mol. Cell* **2**, 887–893 (1998).
44. Wada, T. *et al.* Role of the Src homology 2 (SH2) domain and C-terminus tyrosine phosphorylation sites of SH2-containing inositol phosphatase (SHIP) in the regulation of insulin-induced mitogenesis. *Endocrinology* **140**, 4585–4594 (1999).
45. Clement, S. *et al.* The lipid phosphatase SHIP2 controls insulin sensitivity. *Nature* **409**, 92–97 (2001).
46. Ribon, V. & Saltiel, A. R. Insulin stimulates tyrosine phosphorylation of the proto-oncogene product of c-Cbl in 3T3-L1 adipocytes. *Biochem. J.* **324**, 839–845 (1997).
47. Ribon, V., Herrera, R., Kay, B. K. & Saltiel, A. R. A role for CAP, a novel, multifunctional Src homology 3 domain-containing protein in formation of actin stress fibers and focal adhesions. *J. Biol. Chem.* **273**, 4073–4080 (1998).
48. Ribon, V., Johnson, J. H., Camp, H. S. & Saltiel, A. R. Thiazolidinediones and insulin resistance: peroxisome proliferator activated receptor gamma activation stimulates expression of the CAP gene. *Proc. Natl Acad. Sci. USA* **95**, 14751–14756 (1998).
49. Kimura, A., Baumann, C. A., Chiang, S. H. & Saltiel, A. R. The sorbin homology domain: a motif for the targeting of proteins to lipid rafts. *Proc. Natl Acad. Sci. USA* **98**, 9098–9103 (2001).
50. Baumann, C. A. *et al.* CAP defines a second signalling pathway required for insulin-stimulated glucose transport. *Nature* **407**, 202–207 (2000).
51. Chiang, S. H. *et al.* Insulin-stimulated GLUT4 translocation requires the CAP-dependent activation of TC10. *Nature* **410**, 944–948 (2001).
52. Watson, R. T. *et al.* Lipid raft microdomain compartmentalization of TC10 is required for insulin signaling and GLUT4 translocation. *J. Cell Biol.* **154**, 829–840 (2001).
53. Boulton, T. G. *et al.* ERKs: a family of protein-serine/threonine kinases that are activated and tyrosine phosphorylated in response to insulin and NGF. *Cell* **65**, 663–675 (1991).
54. Lazar, D. F. *et al.* Mitogen-activated protein kinase kinase inhibition does not block the stimulation of glucose utilization by insulin. *J. Biol. Chem.* **270**, 20801–20807 (1995).
55. Raught, B., Gingras, A. C. & Sonenberg, N. The target of rapamycin (TOR) proteins. *Proc. Natl Acad. Sci. USA* **98**, 7037–7044 (2001).
56. Nave, B. T., Ouwens, M., Withers, D. J., Alessi, D. R. & Shepherd, P. R. Mammalian target of rapamycin is a direct target for protein kinase B: identification of a convergence point for opposing effects of insulin and amino-acid deficiency on protein translation. *Biochem. J.* **344**, 427–431 (1999).
57. Thomas, G. & Hall, M. N. TOR signalling and control of cell growth. *Curr. Opin. Cell Biol.* **9**, 782–787 (1997).
58. Miron, M. *et al.* The translational inhibitor 4E-BP is an effector of PI(3)K/Akt signalling and cell growth in *Drosophila*. *Nature Cell Biol.* **3**, 596–601 (2001).
59. Cross, D. A., Alessi, D. R., Cohen, P., Andjelkovic, M. & Hemmings, B. A. Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase B. *Nature* **378**, 785–789 (1995).
60. Brady, M. J., Nairn, A. C. & Saltiel, A. R. The regulation of glycogen synthase by protein phosphatase 1 in 3T3-L1 adipocytes. Evidence for a potential role for DARPP-32 in insulin action. *J. Biol. Chem.* **272**, 29698–29703 (1997).
61. Newgard, C. B., Brady, M. J., O'Doherty, R. M. & Saltiel, A. R. Organizing glucose disposal: emerging roles of the glycogen targeting subunits of protein phosphatase-1. *Diabetes* **49**, 1967–1977 (2000).
62. Michael, M. D. *et al.* Loss of insulin signaling in hepatocytes leads to severe insulin resistance and progressive hepatic dysfunction. *Mol. Cell* **6**, 87–97 (2000).
63. Bergman, R. N. & Ader, M. Free fatty acids and pathogenesis of type 2 diabetes mellitus. *Trends Endocrinol. Metab.* **11**, 351–356 (2000).
64. Bergman, R. N. New concepts in extracellular signaling for insulin action: the single gateway hypothesis. *Recent Prog. Horm. Res.* **52**, 359–385 (1997).
65. Pilkis, S. J. & Granner, D. K. Molecular physiology of the regulation of hepatic gluconeogenesis and glycolysis. *Annu. Rev. Physiol.* **54**, 885–909 (1992).
66. Sutherland, C., O'Brien, R. M. & Granner, D. K. New connections in the regulation of PEPCK gene expression by insulin. *Phil. Trans. R. Soc. Lond. B* **351**, 191–199 (1996).
67. Yoon, J. C. *et al.* Control of hepatic gluconeogenesis through the transcriptional coactivator PGC-1. *Nature* **413**, 131–138 (2001).
68. Shimomura, I. *et al.* Insulin selectively increases SREBP-1c mRNA in the livers of rats with streptozotocin-induced diabetes. *Proc. Natl Acad. Sci. USA* **96**, 13656–13661 (1999).
69. Foretz, M. *et al.* ADD1/SREBP-1c is required in the activation of hepatic lipogenic gene expression by glucose. *Mol. Cell. Biol.* **19**, 3760–3768 (1999).
70. Shimomura, I. *et al.* Decreased IRS-2 and increased SREBP-1c lead to mixed insulin resistance and sensitivity in livers of lipodystrophic and ob/ob mice. *Mol. Cell* **6**, 77–86 (2000).
71. Anthonen, M. W., Ronnstrand, L., Wernstedt, C., Degerman, E. & Holm, C. Identification of novel phosphorylation sites in hormone-sensitive lipase that are phosphorylated in response to isoproterenol and govern activation properties in vitro. *J. Biol. Chem.* **273**, 215–221 (1998).
72. Kitamura, T. *et al.* Insulin-induced phosphorylation and activation of cyclic nucleotide phosphodiesterase 3B by the serine-threonine kinase Akt. *Mol. Cell. Biol.* **19**, 6286–6296 (1999).
73. Cusi, K. *et al.* Insulin resistance differentially affects the PI 3-kinase- and MAP kinase-mediated signaling in human muscle. *J. Clin. Invest.* **105**, 311–320 (2000).
74. Taylor, S. I. & Arioglu, E. Syndromes associated with insulin resistance and acanthosis nigricans. *J. Basic Clin. Physiol. Pharmacol.* **9**, 419–439 (1998).
75. Stern, M. P. Strategies and prospects for finding insulin resistance genes. *J. Clin. Invest.* **106**, 323–327 (2000).
76. Bruning, J. C. *et al.* Development of a novel polygenic model of NIDDM in mice heterozygous for IR and IRS-1 null alleles. *Cell* **88**, 561–572 (1997).
77. Terauchi, Y. *et al.* Development of non-insulin-dependent diabetes mellitus in the double knockout mice with disruption of insulin receptor substrate-1 and β cell glucokinase genes. Genetic reconstitution of diabetes as a polygenic disease. *J. Clin. Invest.* **99**, 861–866 (1997).
78. Kulkarni, R. N. *et al.* Altered function of insulin receptor substrate-1-deficient mouse islets and cultured beta-cell lines. *J. Clin. Invest.* **104**, R69–R75 (1999).
79. Bruning, J. C. *et al.* Role of brain insulin receptor in control of body weight and reproduction. *Science* **289**, 2122–2125 (2000).
80. Zisman, A. *et al.* Targeted disruption of the glucose transporter 4 selectively in muscle causes insulin resistance and glucose intolerance. *Nature Med.* **6**, 924–928 (2000).
81. Katz, E. B., Stenbit, A. E., Hatton, K., DePinho, R. & Charron, M. J. Cardiac and adipose tissue abnormalities but not diabetes in mice deficient in GLUT4. *Nature* **377**, 151–155 (1995).
82. Kim, J. K. *et al.* Redistribution of substrates to adipose tissue promotes obesity in mice with selective insulin resistance in muscle. *J. Clin. Invest.* **105**, 1791–1797 (2000).
83. Shulman, G. I. Cellular mechanisms of insulin resistance. *J. Clin. Invest.* **106**, 171–176 (2000).
84. Hwang, J. H., Pan, J. W., Heydari, S., Hetherington, H. P. & Stein, D. T. Regional differences in intramyocellular lipids in humans observed by in vivo ¹H-MR spectroscopic imaging. *J. Appl. Physiol.* **90**, 1267–1274 (2001).
85. Kim, J. K. *et al.* Tissue-specific overexpression of lipoprotein lipase causes tissue-specific insulin resistance. *Proc. Natl Acad. Sci. USA* **98**, 7522–7527 (2001).
86. Ventre, J. *et al.* Targeted disruption of the tumor necrosis factor- α gene: metabolic consequences in obese and nonobese mice. *Diabetes* **46**, 1526–1531 (1997).
87. Ofei, F., Hurel, S., Newkirk, J., Sopwith, M., Taylor, R. Effects of an engineered human anti-TNF- α antibody (SDP571) on insulin sensitivity and glycemic control in patients with NIDDM. *Diabetes* **45**, 881–885 (1996).
88. Halaas, J. L. *et al.* Weight-reducing effects of the plasma protein encoded by the obese gene. *Science* **269**, 543–546 (1995).
89. Shimomura, I., Hammer, R. E., Ikemoto, S., Brown, M. S. & Goldstein, J. L. Leptin reverses insulin resistance and diabetes mellitus in mice with congenital lipodystrophy. *Nature* **401**, 73–76 (1999).
90. Lee, Y. *et al.* Liporegulation in diet-induced obesity. The antisteatotic role of hyperleptinemia. *J. Biol. Chem.* **276**, 5629–5635 (2001).
91. Fruebis, J. *et al.* Proteolytic cleavage product of 30-kDa adipocyte complement-related protein increases fatty acid oxidation in muscle and causes weight loss in mice. *Proc. Natl Acad. Sci. USA* **98**, 2005–2010 (2001).
92. Yamauchi, T. *et al.* The fat-derived hormone adiponectin reverses insulin resistance associated with both lipodystrophy and obesity. *Nature Med.* **7**, 941–946 (2001).
93. Berg, A. H., Combs, T. P., Du, X., Brownlee, M. & Scherer, P. E. The adipocyte-secreted protein Acrp30 enhances hepatic insulin action. *Nature Med.* **7**, 947–953 (2001).
94. Vionnet, N. *et al.* Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. *Am. J. Hum. Genet.* **67**, 1470–1480 (2000).
95. Steppan, C. M. *et al.* A family of tissue-specific resistin-like molecules. *Proc. Natl Acad. Sci. USA* **98**, 502–506 (2001).
96. Nagaev, I. & Smith, U. Insulin resistance and type 2 diabetes are not related to resistin expression in human fat cells or skeletal muscle. *Biochem. Biophys. Res. Commun.* **285**, 561–564 (2001).
97. Accili, D., Nakae, J., Kim, J. J., Park, B. C. & Rother, K. I. Targeted gene mutations define the roles of insulin and IGF-1 receptors in mouse embryonic development. *J. Pediatr. Endocrinol. Metab.* **12**, 475–485 (1999).
98. Postic, C. *et al.* Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic β cell-specific gene knock-outs using Cre recombinase. *J. Biol. Chem.* **274**, 305–315 (1999).
99. Pessin, J. E., Thurmond, D. C., Elmendorf, J. S., Coker, K. J. & Okada, S. Molecular basis of insulin-stimulated GLUT4 vesicle trafficking. Location! Location! Location! *J. Biol. Chem.* **274**, 2593–2596 (1999).
100. Guilherme, A. *et al.* Perinuclear localization and insulin responsiveness of GLUT4 requires cytoskeletal integrity in 3T3-L1 adipocytes. *J. Biol. Chem.* **275**, 38151–38159 (2000).

Acknowledgements

We apologize to the many investigators whose work we could not cite owing to a limit on the number of references.

Division of Clinical Biochemistry, Department of Internal Medicine, University Medical Centre, 1211 Geneva 4, Switzerland
(e-mail: claes.wollheim@medecine.unige.ch; pierre.maechler@medecine.unige.ch)

Mitochondria are present in most eukaryotic cells, varying in number from hundreds to thousands¹, and have also been visualized as a continuous network². Their origin is generally thought to lie in the endosymbiotic association of oxidative bacteria and glycolytic proto-eukaryotic cells³. The phenotype resulting from the absence of mitochondria is illustrated by mammalian peripheral red blood cells, which depend entirely on glycolysis for their energy supply. The endosymbiotic hypothesis of mitochondrial origin is

The mtDNA is transcribed and translated within the mitochondrion. The nuclear genome specifies the remainder (the majority) of the enzyme subunits and the other

Human mitochondrial DNA

0.16,569

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

T3264G

T3271C

G3316A

T3394C

T3398C

T16189C

T14709C

T14577C

C12258A

7472insC

A8296G

A8344G

T8356C

ATP8

ATP6

COX3

COX2

COX1

ND3

ND4L

ND4

ND5

ND6

Cyt b

ND1

ND2

ND3

ND4

ND5

ND6

12S rRNA

16S rRNA

A3243G

C3254A

C3256T

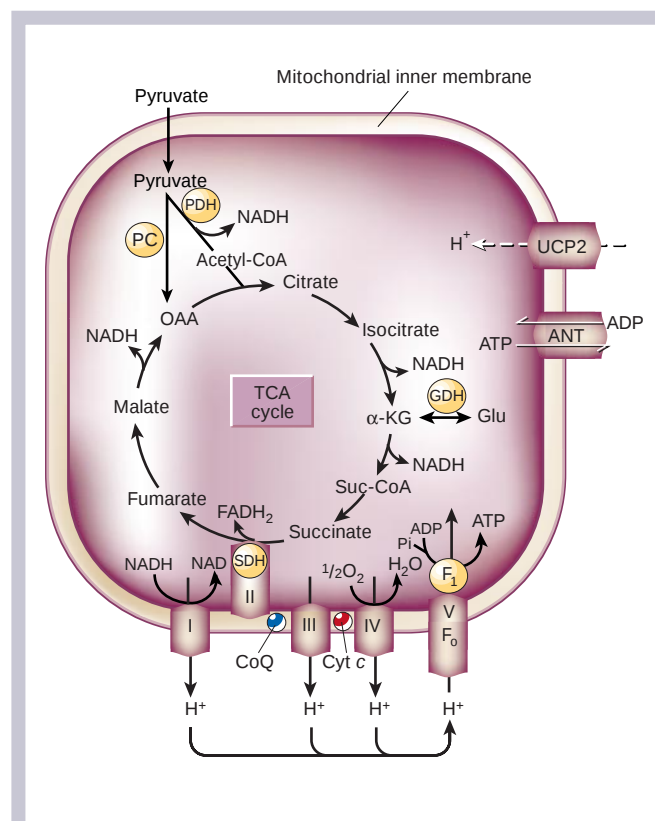


Figure 2 The TCA cycle and respiratory chain in a mitochondrion. Substrate oxidation in the TCA cycle activates the respiratory chain, leading to the generation of ATP, which is subsequently translocated to the cytosol. ANT, adenine nucleotide translocator; GDH, glutamate dehydrogenase; Glu, glutamate; KG, ketoglutarate; OAA, oxaloacetate; PC, pyruvate carboxylase; PDH, pyruvate dehydrogenase; Suc-CoA, succinyl-CoA; SDH, succinate dehydrogenase; TCA, tricarboxylic acid; UCP2, uncoupling protein 2.

mitochondrial proteins; these are synthesized in the cytosol and imported into the mitochondrion⁴. The nucleus also controls the transcriptional activity of mtDNA through regulatory proteins such as mitochondrial transcription factor A (TFAM), which are encoded by nuclear genes⁵. The mtDNA is maternally inherited because of the non-persistence of sperm mtDNA in the zygote after fertilization. The mitochondrial genome exists in multiple copies in every cell. It is highly susceptible to mutation as, in contrast to nuclear DNA, mtDNA consists only of coding sequences and its repair mechanisms are poor. Consequently, it is particularly sensitive to oxidative stress. Moreover, it is juxtaposed to the respiratory chain, which generates mutagenic oxygen derivatives⁶.

The mitochondria are the main source of energy, essentially ATP, which is required for such vital cellular functions as the maintenance of transmembrane ion gradients, protein synthesis and vesicular transport. Additionally, in the pancreatic β -cell, ATP and other mitochondrial factors accomplish the coupling of glucose metabolism to insulin secretion. The mitochondria can be activated by the three classes of fuel: amino acids, fatty acids and carbohydrates, the latter being of most relevance in β -cells under physiological conditions. The principal mitochondrial substrate is pyruvate, formed essentially by glycolysis. Pyruvate carbons enter the tricarboxylic acid cycle (TCA cycle) in the mitochondrial matrix, in which substrates are oxidized with the formation of CO_2 and the reduction of NAD^+ and FADH to NADH and FADH_2 , respectively. These provide electrons to the respiratory chain upon their reoxidation (Fig. 2).

The electron flow along the respiratory chain drives the extrusion of protons from the mitochondrial matrix, which establishes a steep

electrochemical gradient across the inner mitochondrial membrane. The mitochondrial membrane potential is created by complexes I, III and IV of the respiratory chain and is negative inside. ATP synthase (complex V) in the mitochondrial membrane catalyses the condensation of ADP with inorganic phosphate to yield ATP. The generation of this 'high-energy bond' is powered by the diffusion of protons back into the matrix through ATP synthase. Finally, ATP is transferred to the cytosol in exchange for ADP by the adenine nucleotide translocator (ANT). Electrons can enter the respiratory chain at both complexes I (NADH) and II (FADH_2). The latter complex, succinate dehydrogenase, is also an integral part of the TCA cycle. The entire bioenergetic process is regulated not only by substrate flux but also by Ca^{2+} , which increases the activity of several mitochondrial dehydrogenases. An increase in free cytosolic Ca^{2+} that occurs at cell activation is relayed to the mitochondrial matrix by way of a Ca^{2+} uniporter, thus ensuring that the energy requirements of the cell are covered^{7,8}. Ca^{2+} entry is favoured by activation of the respiratory chain, for instance by glucose in the β -cell.

Stimulus–secretion coupling of insulin release

Blood glucose level is tightly controlled by insulin secretion from pancreatic β -cells and insulin action on liver, muscle and other target tissues. The β -cell is poised to adapt insulin secretion to the fluctuations in blood glucose concentration (Fig. 3). Glucose equilibrates across the plasma membrane and its phosphorylation by glucokinase to glucose-6-phosphate determines the rate of glycolysis and the rate of pyruvate generation^{9,10}. Thus, when blood glucose is high, the rate of glycolysis in the β -cell will increase. In the β -cell, pyruvate is the main product of glycolysis, as little lactate is produced¹¹. The supply of cytosolic NAD^+ , necessary to maintain high rates of glycolysis, is therefore ensured by mitochondrial shuttles⁹. Compared to other cell types, an unusually high proportion of glucose-derived carbon enters the mitochondria in the form of pyruvate and enters the TCA cycle.

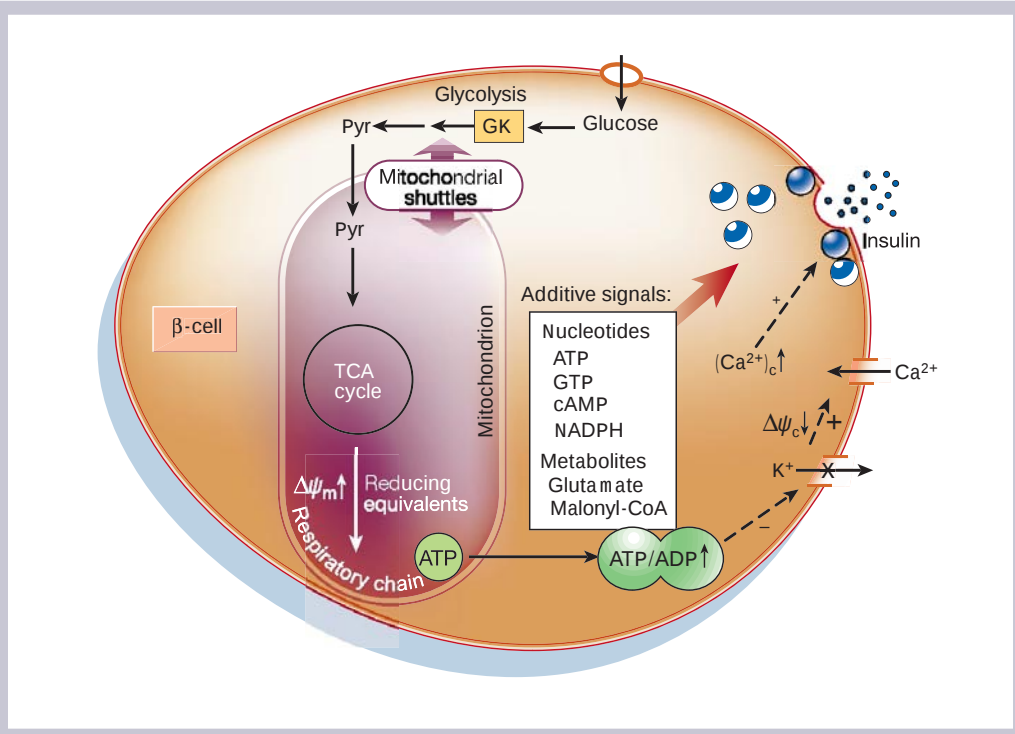
Subsequent oxidative metabolism provides the link between the glucose stimulus and insulin secretion^{12,13}. In the mitochondria, pyruvate is a substrate for both pyruvate dehydrogenase and pyruvate carboxylase, thereby ensuring anaplerosis (provision of carbons) to the TCA cycle. Electron transfer from the TCA cycle to the respiratory chain by NADH and FADH_2 promotes the generation of ATP, which is exported to the cytosol. The increase in the ATP:ADP ratio in the cytosol causes depolarization of the plasma membrane by the closure of ATP-sensitive K^+ channels (K_{ATP})¹⁴. This allows the opening of voltage-sensitive Ca^{2+} channels^{14,15}, similar to those expressed in other excitable cells. This is the key step by which glucose stimulates insulin secretion, as the increase in cytosolic Ca^{2+} is the main trigger for exocytosis, the process by which the insulin-containing secretory granules fuse with the plasma membrane^{15,16}.

The importance of membrane-potential control is illustrated by the syndrome of persistent hyperinsulinaemic hypoglycaemia of infancy (PHHI). It is most frequently caused by mutations in one of the two subunits (the sulphonylurea receptor and KIR6.2) of the K_{ATP} channel, resulting in uncontrolled Ca^{2+} -mediated hypersecretion of insulin¹⁷. However, PHHI patients often retain some glucose-stimulated insulin secretion above the constitutively increased basal rate¹⁸. This supports *in vitro* observations that suggest the existence of a K_{ATP} channel-independent effect of glucose¹⁹. Glucose is thus capable of eliciting a partial secretory response under conditions of clamped, elevated cytosolic Ca^{2+} concentration without affecting the plasma membrane potential. It can be concluded that ATP generated in the mitochondria is the main coupling messenger in insulin secretion, but that other metabolic factors are necessary for the full development of the secretory response.

Signals and messengers for insulin exocytosis

Ultrastructural examination of the β -cell has suggested that the mitochondria are often in close proximity to the secretory insulin

Figure 3 Model for coupling of glucose metabolism to insulin secretion in the β -cell. Glucose is phosphorylated by glucokinase (GK) and converted to pyruvate (Pyr) by glycolysis. Pyruvate preferentially enters the mitochondria and fuels the TCA cycle, resulting in the transfer of reducing equivalents to the respiratory chain, leading to hyperpolarization of the mitochondrial membrane ($\Delta\psi_m \uparrow$) and generation of ATP. Subsequently, closure of K_{ATP} -channels depolarizes the cell membrane ($\Delta\psi_c \downarrow$). This opens voltage-gated Ca^{2+} channels, raising the cytosolic Ca^{2+} concentration ($[Ca^{2+}]_c \uparrow$), which triggers insulin exocytosis. Several putative messengers, or additive signals, proposed to participate in the metabolism–secretion coupling are indicated.



granules (Fig. 4). This may facilitate metabolism–secretion coupling, as ATP is a major permissive factor for movement of insulin granules and for priming of exocytosis^{16,20}. This action of ATP is distinct and complementary to its action on the K_{ATP} channel. Thus, channel activity and exocytosis are both determined by the cytosolic ATP:ADP ratio. Glucose also generates GTP, which could trigger insulin exocytosis through GTPases^{21,22}. GTP is formed in mitochondria by the TCA cycle, but is trapped in the organelle. In the cytosol, GTP is formed mainly through the action of nucleoside diphosphate kinase. In contrast to ATP, GTP is capable of initiating insulin exocytosis in a Ca^{2+} -independent fashion, which qualifies it as a messenger molecule^{22–24}. It is not known whether GTP

acts by way of a monomeric or heterotrimeric G protein that directly controls exocytosis^{16,25}.

Cyclic AMP (cAMP) — the universal second messenger — is generated at the plasma membrane from ATP and potentiates glucose-stimulated insulin secretion. Many neurotransmitters and hormones, including glucagon and the intestinal hormones glucagon-like peptide 1 (GLP-1) and gastric inhibitory polypeptide (GIP), increase cAMP levels in the β -cell by activating adenylyl cyclase²⁶. Although glucose itself is inefficient in stimulating the production of cAMP, permissive levels of cAMP are necessary for normal responsiveness of secretion²⁷. The most important target of cAMP is the exocytotic machinery, where the messenger acts in both a protein kinase A-dependent and independent manner^{22,23,28–30}. Among other putative nucleotide messengers, NADH and NADPH are generated by glucose metabolism⁹. The rise in pyridine nucleotides precedes the increase in cytosolic Ca^{2+} (ref. 31) and their levels rise more rapidly in the cytosol than in the mitochondria³². It remains to be determined whether pyridine nucleotides participate in β -cell activation through effects on ion fluxes or on the process of exocytosis.

In glucose-stimulated β -cells, the TCA cycle intermediate citrate is exported from mitochondria. In the cytosol, citrate carbons are transferred to coenzyme A (CoA) to form malonyl-CoA, which is a lipid precursor. Malonyl-CoA inhibits transport of fatty acids into the mitochondria and their subsequent oxidation, thereby favouring the synthesis of long-chain acyl-CoAs in the cytosol. This metabolic switch led to the proposal that malonyl-CoA acts as a metabolic coupling factor in insulin secretion³³. The long-chain acyl-CoA hypothesis was substantiated by the observation that palmitoyl-CoA enhances Ca^{2+} -evoked insulin exocytosis³⁴. Disruption of malonyl-CoA accumulation during glucose stimulation did not, however, attenuate the secretory response³⁵. Therefore, the role of long-chain acyl-CoA derivatives in metabolism–secretion coupling remains controversial.

Studies in a permeabilized β -cell model have shown a direct link between mitochondrial activation and insulin exocytosis³⁶. Under conditions of permissive Ca^{2+} concentrations, stimulation of mitochondrial metabolism³⁷ indicated the generation of a coupling factor³⁶. The factor was subsequently identified as glutamate³⁸, which can be produced from the TCA cycle intermediate α -ketoglutarate or

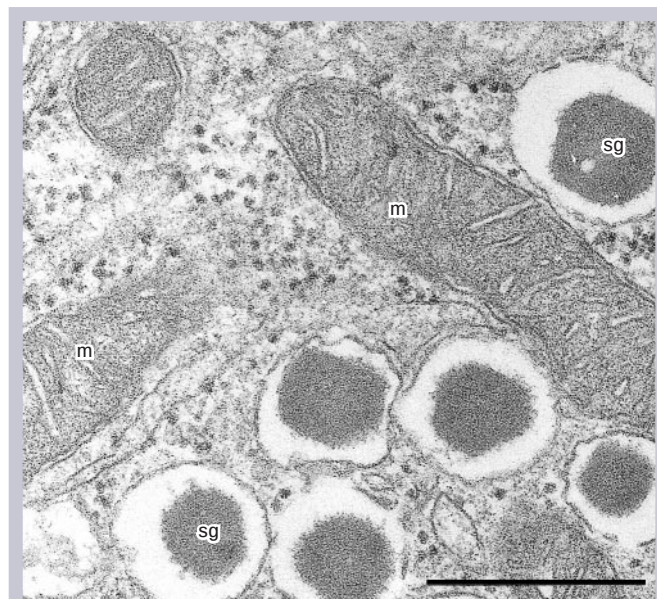


Figure 4 Electron micrograph of part of a rat β -cell showing mitochondria (m) and insulin-containing secretory granules (sg). The scale bar represents 0.5 μ m. Reprinted with permission from ref. 13.

by transamination reactions³⁹. In non-permeabilized cell preparations, a membrane-permeant derivative of glutamate was also shown to enhance the action of glucose in insulin secretion^{38,40}. Further studies demonstrated a positive correlation between cellular glutamate concentrations and the secretory response to glucose⁴¹. At present, the mechanism of glutamate action on exocytosis is unknown. In this scenario the classical neurotransmitter glutamate is allocated a new role, that of an intracellular messenger or cofactor in insulin secretion⁴².

Mitochondrial dysfunction in the β -cell

It was established several decades ago that blockade of the respiratory chain inhibits glucose-stimulated insulin secretion. This conclusion was based on the use of various mitochondrial poisons and lowering of the oxygen supply to the β -cell⁴³. More recently, the activity of the respiratory chain was impaired by suppressing the production of those enzyme subunits encoded by mtDNA, which creates so-called rho⁰ cells⁴⁴. In this way, ethidium bromide and other chemical treatment of β -cell lines resulted in depletion of mtDNA with preservation of insulin biosynthesis and cell viability, albeit with a reduced proliferation rate^{45–48}. In such cells, glucose-induced insulin release is absent, as a result of defective respiratory-chain activation and the loss of mitochondrial ATP production^{45–47}. The defect is of mitochondrial origin, as specific mitochondrial substrates were equally inefficient, whereas the secretory response to Ca²⁺-raising agents was still present^{46–48}. Elegant experiments showed that the secretory response to glucose could be restored by replenishing rho⁰ β -cells with normal mitochondria⁴⁵.

Expression of mtDNA is controlled by a nucleus-encoded transcription factor, TFAM, and disruption of this gene in the mouse is lethal⁵. The phenotype of the heterozygous knockout mouse revealed that the heart is highly sensitive to respiratory-chain deficiency⁵. The β -cell-specific deletion of the *Tfam* gene caused a diabetic phenotype⁴⁹. The islets of these mice showed attenuated respiratory-chain activation and diminished secretory response to glucose. These transgenic animals represent the first model of human mitochondrial diabetes, which, it should be noted, has been linked mostly to mutations in tRNA genes (Fig. 1). Taken together, the *in vitro* and *in vivo* studies highlight the pivotal role of mitochondria in stimulus–secretion coupling in the β -cell.

mtDNA mutations and diabetes

Point mutations or deletions in mtDNA have been associated with a large spectrum of diseases, with symptoms such as muscle weakness, cardiomyopathy, optic nerve atrophy, retinal dystrophy, impaired hearing and hyperglycaemia (diabetes mellitus). Point mutations in the mitochondrial tRNA genes are the primary cause of these pathological manifestations (Fig. 1). A specific, maternally inherited form of diabetes mellitus was first linked to mutations in the mtDNA in 1992 (refs 50, 51). Often associated with neurosensory deafness, it is also called maternally inherited mitochondrial diabetes and deafness (MIDD). The most frequent mutation encountered is the A3,243G mutation in the gene for tRNA^{Leu} (bearing the anticodon UUR)^{51,52}. Altogether, mitochondrial diabetes accounts for approximately 1% of all cases of diabetes⁵³. The mutation reduces the stability of mtDNA-encoded proteins and the binding of leucine to its tRNA, resulting in decreased synthesis of certain mitochondrial proteins⁵³. The same mutation causes the MELAS syndrome (mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes)⁵⁴, which may be associated with diabetes⁵³.

How is it possible that the same A3,243G mutation in mtDNA can give rise to two different phenotypes affecting different tissues? On the one hand, this is due to tissue-specific segregation of the mutant mtDNA copies. In the course of cell division, an increasing proportion of mutant mtDNA substituting for wild-type copies is referred to as heteroplasmy, a process aggravated in post-mitotic cells. On the other hand, the decreased mitochondrial function will

impact differently on cell viability according to tissue sensitivity to apoptosis. The A3,243G mutation has been associated with reduction of islet mass involving both β -cells and the neighbouring glucagon-producing α -cells^{55,56}. In post-mortem studies, diabetic patients with A3,243G mutation showed a degree of heteroplasmy ranging from 32% to 63% (refs 55, 56). It is possible that the extent of heteroplasmy may participate in the lowering of the cellular energy capacity below the bioenergetic threshold¹. The consequence of this would be that the organelle could no longer fulfil the energetic and signalling requirements for glucose-stimulated insulin secretion, a concept also supported by clinical observations⁵⁷. In contrast, secretion induced by the receptor agonist glucagon, acting by way of cAMP effects on exocytosis, was preserved in A3,243G patients⁵⁷. This form of diabetes may or may not require insulin injections, and is usually not associated with resistance of muscle and other tissues to the action of the hormone^{52,53,57}.

The molecular diagnosis of mitochondrial diabetes is complicated by an invariably low degree of heteroplasmy in the peripheral white blood cells usually used for genetic analysis. *Ex vivo* mitochondrial dysfunction associated with the A3,243G mutation was demonstrated in skin fibroblasts⁵⁸ or in cells enriched for patient mutant mitochondria⁵⁹. A similar conclusion was drawn from a study on another mtDNA mutation associated with a neurodegenerative disease⁶⁰. The mitochondrial diabetes phenotype illustrates the importance of normal respiratory-chain function in the β -cell for glucose homeostasis.

Other pathophysiological considerations

In contrast to the rare monogenic mitochondrial diabetes, type 2 or non-insulin-dependent diabetes mellitus is common and polygenic in nature⁶¹. Patients usually display both resistance to insulin at its target tissues (mainly skeletal muscle) as well as defective insulin secretion⁶². Although the contribution of variations in mtDNA to the development of type 2 diabetes is unknown, a 50% decrease in mtDNA copy number in skeletal muscle of type 2 diabetics was observed⁶³. A reduced mtDNA content was also reported in peripheral blood cells in such patients even before the onset of the disease⁶⁴. The corresponding information for β -cells is lacking.

The diabetic state is generally characterized by accelerated tissue ageing, perhaps related to mitochondrial dysfunction. Accumulation of point mutations in mtDNA has been reported to occur in an age-dependent manner in humans⁶⁵. There is an age-related increase in the production of reactive oxygen species (ROS), while concurrently the defence mechanisms against these free radicals are diminishing⁶. The mitochondria are the principal source of ROS resulting from imperfect electron transport. Normally, only 0.1% of total oxygen consumption leaks to ROS generation, but the percentage becomes greater in the ageing tissue⁶. This deleterious process is amplified by the diminishing natural enzymatic defences (for example, catalase and superoxide dismutase). The low expression of these protective enzymes⁶⁶ makes the β -cell particularly susceptible to ROS action⁶⁷. In addition to their acute effects, ROS may also lead to increased mutation in mtDNA, exacerbated by the limited repair capacity. These considerations have prompted the suggestion that ROS may participate in the impairment of glucose-induced insulin secretion seen both in ageing and in type 2 diabetes⁶⁸.

Further clues on β -cell function come from other forms of monogenic diabetes. Different subclasses of maturity-onset diabetes of the young (MODY) represent such monogenic forms of diabetes with autosomal dominant transmission. They are characterized by β -cell dysfunction as a result of mutations in nuclear genes⁶⁹. MODY1 and MODY3 have been linked to mutations in the transcription factors hepatocyte nuclear factor HNF-4 α and HNF-1 α , respectively⁶⁹. MODY3 is the most common form of this inherited disease and explains 2–5% of diabetic cases. Suppression of the *HNF-1 α* gene in mice results in diabetes and impairment of glucose-induced insulin secretion assessed *in vitro*⁷⁰. In cellular model systems, the molecular

basis of the defect has been attributed to deranged mitochondrial metabolism^{71,72}. In particular, the defective respiratory-chain activation correlated with downregulation of the TCA cycle enzyme α -ketoglutarate dehydrogenase, accompanied by an upregulation of uncoupling protein 2 (UCP2)⁷².

UCP2 is an inner mitochondrial membrane protein that tends to diminish the proton gradient generated by the respiratory chain. Its overexpression in β -cells attenuates ATP generation and insulin secretion in response to glucose⁷³. It is of interest that deletion of the UCP2 gene in mice enhances islet ATP generation and insulin secretion during glucose stimulation⁷⁴. In obese, diabetic hyperlipidaemic rodent models, UCP2 levels in islets were reported to be either lower⁷⁵ or higher⁷⁴ than lean controls. Thus, there is no simple relationship *in vivo* between circulating lipids and UCP2 function. Type 2 diabetics usually have both hyperglycaemia and hyperlipidaemia. This is thought to induce the phenomenon of 'glucolipotoxicity' in the β -cell, leading to lipid accumulation, impaired glucose metabolism and alterations in mitochondria^{76,77}. Chronic exposure of β -cells to fatty acids induces changes in the expression of numerous genes⁷⁶. Among them, UCP2 is induced, which correlates with reduced glucose-evoked insulin secretion⁷⁸. This may be part of an adaptive mechanism protecting the β -cell against oxidants, as indicated by *in vitro* experiments⁷⁹. Elucidation of the adaptation of the mitochondrial machinery is complicated by the multiple influences on the β -cell in the course of the development of the diabetic state.

Possible therapeutic interventions and perspectives

Patients with mitochondrial diabetes or with one of the MODY subforms are treated like any other case of type 2 diabetes. Treatment begins with diet, but usually needs to be supplemented with oral hypoglycaemic agents, in particular the sulphonylureas. Eventually, blood glucose control may require insulin injections.

Mitochondrially targeted therapy of the insulin secretory defect in a rat model of type 2 diabetes has been proposed. Dimethylsuccinate, a precursor of the TCA cycle intermediate succinate, was found to improve the secretory response⁸⁰. Diabetic patients with a mitochondrial DNA mutation have been given long-term treatment with coenzyme Q10, a component of the respiratory chain⁸¹. This resulted in improved insulin secretion but, disappointingly, did not affect diabetic complications (nephropathy, retinopathy and neuropathy). More efficient treatments should be envisaged for such patients. The ultimate goal is the replacement of mutated DNA with normal mtDNA by either gene or cell therapy. Despite much research, these techniques are not yet available. It was, however, shown recently that complementation of normal mtDNA in mice carrying mutated mtDNA restores mitochondrial function⁸². This may open new perspectives for gene therapy of mitochondrial diseases.

In view of the requirement for optimally functioning mitochondria, measures directed to protecting these organelles should be envisaged in diabetes prevention. Even after the disease is manifest, such therapy could preserve partial β -cell sensitivity to glucose. Further definition of the molecular mechanisms underlying the role of mitochondria in cell activation will help to target interventions in diabetes and other metabolic diseases. □

- Wallace, D. C. Mitochondrial diseases in man and mouse. *Science* **283**, 1482–1488 (1999).
- Rizzuto, R. *et al.* Close contacts with the endoplasmic reticulum as determinants of mitochondrial Ca^{2+} responses. *Science* **280**, 1763–1766 (1998).
- Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* **283**, 1476–1481 (1999).
- Neupert, W. Protein import into mitochondria. *Annu. Rev. Biochem.* **66**, 863–917 (1997).
- Larsson, N. G. *et al.* Mitochondrial transcription factor A is necessary for mtDNA maintenance and embryogenesis in mice. *Nature Genet.* **18**, 231–236 (1998).
- Beckman, K. B. & Ames, B. N. The free radical theory of aging matures. *Physiol. Rev.* **78**, 547–581 (1998).
- McCormack, J. G., Halestrap, A. P. & Denton, R. M. Role of calcium ions in regulation of mammalian intramitochondrial metabolism. *Physiol. Rev.* **70**, 391–425 (1990).
- Duchen, M. R. Contributions of mitochondria to animal physiology: from homeostatic sensor to calcium signalling and cell death. *J. Physiol.* **516**, 1–17 (1999).
- Newgard, C. B. & McGarry, J. D. Metabolic coupling factors in pancreatic beta-cell signal transduction. *Annu. Rev. Biochem.* **64**, 689–719 (1995).

- Matschinsky, F. M. A lesson in metabolic regulation inspired by the glucokinase glucose sensor paradigm. *Diabetes* **45**, 223–241 (1996).
- Ishihara, H., Wang, H., Drewes, L. R. & Wollheim, C. B. Overexpression of monocarboxylate transporter and lactate dehydrogenase alters insulin secretory responses to pyruvate and lactate in β cells. *J. Clin. Invest.* **104**, 1621–1629 (1999).
- Schuit, F. *et al.* Metabolic fate of glucose in purified islet cells. Glucose-regulated anaplerosis in beta cells. *J. Biol. Chem.* **272**, 18572–18579 (1997).
- Wollheim, C. B. Beta-cell mitochondria in the regulation of insulin secretion: a new culprit in Type II diabetes. *Diabetologia* **43**, 265–277 (2000).
- Ashcroft, F. M. *et al.* Stimulus-secretion coupling in pancreatic beta cells. *J. Cell Biochem.* **55**, 54–65 (1994).
- Rorsman, P. The pancreatic beta-cell as a fuel sensor: an electrophysiologist's viewpoint. *Diabetologia* **40**, 487–495 (1997).
- Lang, J. Molecular mechanisms and regulation of insulin exocytosis as a paradigm of endocrine secretion. *Eur. J. Endocrinol.* **259**, 3–17 (1999).
- Dunne, M. J. *et al.* Familial persistent hyperinsulinemic hypoglycemia of infancy and mutations in the sulfonylurea receptor. *N. Engl. J. Med.* **336**, 703–706 (1997).
- Grimberg, A. *et al.* Dysregulation of insulin secretion in children with congenital hyperinsulinism due to sulfonylurea receptor mutations. *Diabetes* **50**, 322–328 (2001).
- Henquin, J. C. Triggering and amplifying pathways of regulation of insulin secretion by glucose. *Diabetes* **49**, 1751–1760 (2000).
- Eliasson, L., Renstrom, E., Ding, W. G., Proks, P. & Rorsman, P. Rapid ATP-dependent priming of secretory granules precedes Ca^{2+} -induced exocytosis in mouse pancreatic B-cells. *J. Physiol.* **503**, 399–412 (1997).
- Detimary, P., Van den Bergh, G. & Henquin, J. C. Concentration dependence and time course of the effects of glucose on adenine and guanine nucleotides in mouse pancreatic islets. *J. Biol. Chem.* **271**, 20559–20565 (1996).
- Wollheim, C. B., Ullrich, S., Meda, P. & Vallar, L. Regulation of exocytosis in electrically permeabilized insulin-secreting cells. Evidence for Ca^{2+} dependent and independent secretion. *Biosci. Rept.* **7**, 443–454 (1987).
- Vallar, L., Biden, T. J. & Wollheim, C. B. Guanine nucleotides induce Ca^{2+} -independent insulin secretion from permeabilized RINm5F cells. *J. Biol. Chem.* **262**, 5049–5056 (1987).
- Proks, P., Eliasson, L., Ammala, C., Rorsman, P. & Ashcroft, F. M. Ca^{2+} - and GTP-dependent exocytosis in mouse pancreatic beta-cells involves both common and distinct steps. *J. Physiol.* **496**, 255–264 (1996).
- Iezzi, M., Regazzi, R. & Wollheim, C. B. The Rab3-interacting molecule RIM is expressed in pancreatic beta-cells and is implicated in insulin exocytosis. *FEBS Lett.* **474**, 66–70 (2000).
- Schuit, F. C., Huypens, P., Heimberg, H. & Pipeleers, D. G. Glucose sensing in pancreatic beta-cells: a model for the study of other glucose-regulated cells in gut, pancreas, and hypothalamus. *Diabetes* **50**, 1–11 (2001).
- Huypens, P., Ling, Z., Pipeleers, D. & Schuit, F. Glucagon receptors on human islet cells contribute to glucose competence of insulin release. *Diabetologia* **43**, 1012–1019 (2000).
- Jones, P. M., Fyles, J. M. & Howell, S. L. Regulation of insulin secretion by cAMP in rat islets of Langerhans permeabilized by high-voltage discharge. *FEBS Lett.* **205**, 205–209 (1986).
- Ammala, C. *et al.* Activation of protein kinases and inhibition of protein phosphatases play a central role in the regulation of exocytosis in mouse pancreatic beta cells. *Proc. Natl Acad. Sci. USA* **91**, 4343–4347 (1994).
- Ozaki, N. *et al.* cAMP-GEFII is a direct target of cAMP in regulated exocytosis. *Nature Cell Biol.* **2**, 805–811 (2000).
- Pralong, W. F., Bartley, C. & Wollheim, C. B. Single islet beta-cell stimulation by nutrients: relationship between pyridine nucleotides, cytosolic Ca^{2+} and secretion. *EMBO J.* **9**, 53–60 (1990).
- Patterson, G. H., Knobel, S. M., Arkhammar, P., Thastrup, O. & Piston, D. W. Separation of the glucose-stimulated cytoplasmic and mitochondrial NAD(P)H responses in pancreatic islet beta cells. *Proc. Natl Acad. Sci. USA* **97**, 5203–5207 (2000).
- Prentki, M. *et al.* Malonyl-CoA and long chain acyl-CoA esters as metabolic coupling factors in nutrient-induced insulin secretion. *J. Biol. Chem.* **267**, 5802–5810 (1992).
- Deeney, J. T. *et al.* Acute stimulation with long chain acyl-CoA enhances exocytosis in insulin-secreting cells (HIT T-15 and NMRI beta-cells). *J. Biol. Chem.* **275**, 9363–9368 (2000).
- Antinozzi, P. A., Segall, L., Prentki, M., McGarry, J. D. & Newgard, C. B. Molecular or pharmacologic perturbation of the link between glucose and lipid metabolism is without effect on glucose-stimulated insulin secretion. A re-evaluation of the long-chain acyl-CoA hypothesis. *J. Biol. Chem.* **273**, 16146–16154 (1998).
- Maechler, P., Kennedy, E. D., Pozzan, T. & Wollheim, C. B. Mitochondrial activation directly triggers the exocytosis of insulin in permeabilized pancreatic β -cells. *EMBO J.* **16**, 3833–3841 (1997).
- Maechler, P., Kennedy, E. D., Wang, H. & Wollheim, C. B. Desensitization of mitochondrial Ca^{2+} and insulin secretion responses in the beta cell. *J. Biol. Chem.* **273**, 20770–20778 (1998).
- Maechler, P. & Wollheim, C. B. Glutamate acts as a mitochondrially derived messenger in glucose-induced insulin exocytosis. *Nature* **402**, 685–689 (1999).
- Nissim, I. Newer aspects of glutamine/glutamate metabolism: the role of acute pH changes. *Am. J. Physiol.* **277**, F493–F497 (1999).
- Sener, A. *et al.* Insulinotropic action of glutamic acid dimethyl ester. *Am. J. Physiol.* **267**, E573–E584 (1994).
- Rubi, B., Ishihara, H., Hegardt, F. G., Wollheim, C. B. & Maechler, P. GAD65-mediated glutamate decarboxylation reduces glucose-stimulated insulin secretion in pancreatic beta cells. *J. Biol. Chem.* **276**, 36391–36396 (2001).
- Maechler, P. & Wollheim, C. B. Mitochondrial signals in glucose-stimulated insulin secretion in the beta cell. *J. Physiol.* **529**, 49–56 (2000).
- Malaisse, W. J. *et al.* The stimulus-secretion coupling of glucose-induced insulin release. XXXV. The links between metabolic and cationic events. *Diabetologia* **16**, 331–341 (1979).
- King, M. P. & Attardi, G. Human cells lacking mtDNA: repopulation with exogenous mitochondrial by complementation. *Science* **246**, 500–503 (1989).
- Soejima, A. *et al.* Mitochondrial DNA is required for regulation of glucose-stimulated insulin secretion in a mouse pancreatic beta cell line, MIN6. *J. Biol. Chem.* **271**, 26194–26199 (1996).
- Kennedy, E. D., Maechler, P. & Wollheim, C. B. Effects of depletion of mitochondrial DNA in metabolism secretion coupling in INS-1 cells. *Diabetes* **47**, 374–380 (1998).
- Tsuruzoe, K. *et al.* Creation and characterization of a mitochondrial DNA-depleted pancreatic beta-cell line: impaired insulin secretion induced by glucose, leucine, and sulfonylureas. *Diabetes* **47**, 621–631 (1998).

48. Hayakawa, T. *et al.* Ethidium bromide-induced inhibition of mitochondrial gene transcription suppresses glucose-stimulated insulin release in the mouse pancreatic β -cell line β HC9. *J. Biol. Chem.* **273**, 20300–20307 (1998).
49. Silva, J. P. *et al.* Impaired insulin secretion and β -cell loss in tissue-specific knockout mice with mitochondrial diabetes. *Nature Genet.* **26**, 336–340 (2000).
50. Ballinger, S. W. *et al.* Maternally transmitted diabetes and deafness associated with a 10.4 kb mitochondrial DNA deletion. *Nature Genet.* **1**, 11–15 (1992).
51. van den Ouweland, J. M. *et al.* Mutation in mitochondrial tRNA^{Leu(UUR)} gene in a large pedigree with maternally transmitted type II diabetes mellitus and deafness. *Nature Genet.* **1**, 368–371 (1992).
52. Kadowaki, T. *et al.* A subtype of diabetes mellitus associated with a mutation of mitochondrial DNA. *N. Engl. J. Med.* **330**, 962–968 (1994).
53. Maassen, J. A., van Essen, E., van den Ouweland, J. M. & Lemkes, H. H. Molecular and clinical aspects of mitochondrial diabetes mellitus. *Exp. Clin. Endocrinol. Diabetes* **109**, 127–134 (2001).
54. Goto, Y., Nonaka, I. & Horai, S. A mutation in the tRNA^{Leu(UUR)} gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature* **348**, 651–653 (1990).
55. Kobayashi, T. *et al.* In situ characterization of islets in diabetes with a mitochondrial DNA mutation at nucleotide position 3243. *Diabetes* **46**, 1567–1571 (1997).
56. Otake, S. *et al.* Molecular and histological evaluation of pancreata from patients with a mitochondrial gene mutation associated with impaired insulin secretion. *Biochem. Biophys. Res. Commun.* **259**, 149–156 (1999).
57. Suzuki, Y. *et al.* Diabetes mellitus associated with the 3243 mitochondrial tRNA(Leu)(UUR) mutation: insulin secretion and sensitivity. *Metabolism* **46**, 1019–1023 (1997).
58. James, A. M., Wei, Y. H., Pang, C. Y. & Murphy, M. P. Altered mitochondrial function in fibroblasts containing MELAS or MERRF mitochondrial DNA mutations. *Biochem. J.* **318**, 401–407 (1996).
59. van den Ouweland, J. M., Maechler, P., Wollheim, C. B., Attardi, G. & Maassen, J. A. Functional and morphological abnormalities of mitochondria harbouring the tRNA(Leu)(UUR) mutation in mitochondrial DNA derived from patients with maternally inherited diabetes and deafness (MIDD) and progressive kidney disease. *Diabetologia* **42**, 485–492 (1999).
60. Brini, M. *et al.* A calcium signaling defect in the pathogenesis of a mitochondrial DNA inherited oxidative phosphorylation deficiency. *Nature Med.* **5**, 951–954 (1999).
61. Froguel, P. & Velho, G. Genetic determinants of type 2 diabetes. *Recent Prog. Horm. Res.* **56**, 91–105 (2001).
62. Polonsky, K. S., Sturis, J. & Bell, G. I. Non-insulin-dependent diabetes mellitus—a genetically programmed failure of the beta cell to compensate for insulin resistance. *N. Engl. J. Med.* **334**, 777–783 (1996).
63. Antonetti, D. A., Reynet, C. & Kahn, C. R. Increased expression of mitochondrial-encoded genes in skeletal muscle of humans with diabetes mellitus. *J. Clin. Invest.* **95**, 1383–1388 (1995).
64. Lee, H. K. *et al.* Decreased mitochondrial DNA content in peripheral blood precedes the development of non-insulin-dependent diabetes mellitus. *Diabetes Res. Clin. Pract.* **42**, 161–167 (1998).
65. Michikawa, Y., Mazzucchelli, F., Bresolin, N., Scarlato, G. & Attardi, G. Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. *Science* **286**, 774–779 (1999).
66. Tiedge, M., Lortz, S., Drinkgern, J. & Lenzen, S. Relation between antioxidant enzyme gene expression and antioxidative defense status of insulin-producing cells. *Diabetes* **46**, 1733–1742 (1997).
67. Maechler, P., Jornot, L. & Wollheim, C. B. Hydrogen peroxide alters mitochondrial activation and insulin secretion in pancreatic beta cells. *J. Biol. Chem.* **274**, 27905–27913 (1999).
68. Coordt, M. C., Ruhe, R. C. & McDonald, R. B. Aging and insulin secretion. *Proc. Soc. Exp. Biol. Med.* **209**, 213–222 (1995).
69. Hattersley, A. T. Maturity-onset diabetes of the young: clinical heterogeneity explained by genetic heterogeneity. *Diabet. Med.* **15**, 15–24 (1998).
70. Pontoglio, M. *et al.* Defective insulin secretion in hepatocyte nuclear factor 1 α -deficient mice. *J. Clin. Invest.* **101**, 2215–2222 (1998).
71. Wang, H., Maechler, P., Hagenfeldt, K. A. & Wollheim, C. B. Dominant-negative suppression of HNF-1 α function results in defective insulin gene transcription and impaired metabolism-secretion coupling in a pancreatic β -cell line. *EMBO J.* **17**, 6701–6713 (1998).
72. Wang, H., Antinozzi, P. A., Hagenfeldt, K. A., Maechler, P. & Wollheim, C. B. Molecular targets of a human HNF1 α mutation responsible for pancreatic β -cell dysfunction. *EMBO J.* **19**, 4257–4264 (2000).
73. Chan, C. B. *et al.* Increased uncoupling protein-2 levels in beta-cells are associated with impaired glucose-stimulated insulin secretion: mechanism of action. *Diabetes* **50**, 1302–1310 (2001).
74. Zhang, C. Y. *et al.* Uncoupling protein-2 negatively regulates insulin secretion and is a major link between obesity, beta cell dysfunction, and type 2 diabetes. *Cell* **105**, 745–755 (2001).
75. Wang, M. Y. *et al.* Adenovirus-mediated overexpression of uncoupling protein-2 in pancreatic islets of Zucker diabetic rats increases oxidative activity and improves beta-cell function. *Diabetes* **48**, 1020–1025 (1999).
76. Unger, R. H., Zhou, Y. T. & Orci, L. Regulation of fatty acid homeostasis in cells: novel role of leptin. *Proc. Natl Acad. Sci. USA* **96**, 2327–2332 (1999).
77. Roduit, R. *et al.* Glucose down-regulates the expression of the peroxisome proliferator-activated receptor- α gene in the pancreatic beta-cell. *J. Biol. Chem.* **275**, 35799–35806 (2000).
78. Lameloise, N., Muzzin, P., Prentki, M. & Assimacopoulos-Jeannet, F. Uncoupling protein 2: a possible link between fatty acid excess and impaired glucose-induced insulin secretion? *Diabetes* **50**, 803–809 (2001).
79. Li, L. X., Skorpen, F., Egeberg, K., Jorgensen, I. H. & Grill, V. Uncoupling protein-2 participates in cellular defense against oxidative stress in clonal beta-cells. *Biochem. Biophys. Res. Commun.* **282**, 273–277 (2001).
80. Garcia-Martinez, J. A., Cancelas, J., Villanueva-Penacarrillo, M. L., Valverde, I. & Malaisse, W. J. Prolongation of the insulinotropic action of glucagon-like peptide 1 by the dimethyl ester of succinic acid in an animal model of type-2 diabetes. *Int. J. Mol. Med.* **6**, 319–321 (2000).
81. Suzuki, S. *et al.* The effects of coenzyme Q10 treatment on maternally inherited diabetes mellitus and deafness, and mitochondrial DNA 3243 (A to G) mutation. *Diabetologia* **41**, 584–588 (1998).
82. Nakada, K. *et al.* Inter-mitochondrial complementation: mitochondria-specific system preventing mice from expression of disease phenotypes by mutant mtDNA. *Nature Med.* **7**, 934–940 (2001).

Acknowledgements

We apologize to colleagues whose papers were not cited owing to space limitations. We are grateful to L. Orci for kindly providing the Fig. 4, to T. Pozzan and P. Antinozzi for most helpful discussions and to the Swiss National Science Foundation for continued support of our research.

Biochemistry and molecular cell biology of diabetic complications

Michael Brownlee

Departments of Medicine and Pathology, and Diabetes Research and Training Center, Albert Einstein College of Medicine, Bronx, New York 10461, USA (e-mail: brownlee@aeom.yu.edu)

Diabetes-specific microvascular disease is a leading cause of blindness, renal failure and nerve damage, and diabetes-accelerated atherosclerosis leads to increased risk of myocardial infarction, stroke and limb amputation. Four main molecular mechanisms have been implicated in glucose-mediated vascular damage. All seem to reflect a single hyperglycaemia-induced process of overproduction of superoxide by the mitochondrial electron-transport chain. This integrating paradigm provides a new conceptual framework for future research and drug discovery.

All forms of diabetes are characterized by chronic hyperglycaemia and the development of diabetes-specific microvascular pathology in the retina, renal glomerulus and peripheral nerve. As a consequence of its microvascular pathology, diabetes is a leading cause of blindness, end-stage renal disease and a variety of debilitating neuropathies. Diabetes is also associated with accelerated atherosclerotic macrovascular disease affecting arteries that supply the heart, brain and lower extremities. As a result, patients with diabetes have a much higher risk of myocardial infarction, stroke and limb amputation. Large prospective clinical studies show a strong relationship between glycaemia and diabetic microvascular complications in both type 1 and type 2 diabetes^{1,2}. Hyperglycaemia and insulin resistance both seem to have important roles in the pathogenesis of macrovascular complications²⁻⁵.

Diabetes-specific microvascular disease in the retina, glomerulus and vasa nervorum has similar pathophysiological features. Early in the course of diabetes, intracellular hyperglycaemia causes abnormalities in blood flow and increased vascular permeability. This reflects decreased activity of vasodilators such as nitric oxide, increased activity of vasoconstrictors such as angiotensin II and endothelin-1, and elaboration of permeability factors such as vascular endothelial growth factor (VEGF). Quantitative and qualitative abnormalities of extracellular matrix contribute to an irreversible increase in vascular permeability. With time, microvascular cell loss occurs, in part as a result of programmed cell death, and there is progressive capillary occlusion due both to extracellular matrix overproduction induced by growth factors such as transforming growth factor- β (TGF- β), and to deposition of extravasated periodic acid-Schiff-positive plasma proteins. Hyperglycaemia may also decrease production of trophic factors for endothelial and neuronal cells. Together, these changes lead to oedema, ischaemia and hypoxia-induced neovascularization in the retina, proteinuria, mesangial matrix expansion and glomerulosclerosis in the kidney, and multifocal axonal degeneration in peripheral nerves.

The pathogenesis of atherosclerosis in non-diabetics has been extensively described in recent reviews, and begins with endothelial dysfunction⁶. In diabetic arteries, endothelial dysfunction seems to involve both insulin resistance specific to the phosphatidylinositol-3-OH kinase pathway^{7,8}

and hyperglycaemia. Pathway-selective insulin resistance results in decreased endothelial production of the anti-atherogenic molecule nitric oxide, and increased potentiation of proliferation of vascular smooth muscle cells and production of plasminogen activator inhibitor-1 (PAI-1) via the Ras $\rightarrow$ Raf $\rightarrow$ MEK kinase $\rightarrow$ mitogen-activated protein (MAP) kinase pathway⁷. Hyperglycaemia itself also inhibits production of nitric oxide in arterial endothelial cells⁹ and stimulates production of PAI-1 (ref. 10).

Both insulin resistance and hyperglycaemia have also been implicated in the pathogenesis of diabetic dyslipidaemia. The role of insulin resistance has been reviewed recently⁵. Hyperglycaemia seems to cause raised levels of atherogenic cholesterol-enriched apolipoprotein B-containing remnant particles by reducing expression of the heparan sulphate proteoglycan perlecan on hepatocytes⁴. Associations of atherosclerosis and atherosclerosis risk factors with glycaemia have been shown over a broad range of glucose tolerance, from normal to diabetic. Postprandial hyperglycaemia may be more predictive of atherosclerosis than is fasting plasma glucose level or haemoglobin A1c¹¹. Whether postprandial hyperglycaemia is an independent risk factor is controversial and requires further study.

Mechanisms of hyperglycaemia-induced damage

How do these diverse microvascular and macrovascular pathologies all result from hyperglycaemia? Four main hypotheses about how hyperglycaemia causes diabetic complications have generated a large amount of data, as well as several clinical trials based on specific inhibitors of these mechanisms. The four hypotheses are: increased polyol pathway flux; increased advanced glycation end-product (AGE) formation; activation of protein kinase C (PKC) isoforms; and increased hexosamine pathway flux. Until recently there was no unifying hypothesis linking these four mechanisms.

Increased polyol pathway flux

Aldose reductase (alditol:NAD(P)⁺ 1-oxidoreductase, EC 1.1.1.21) is the first enzyme in the polyol pathway. It is a cytosolic, monomeric oxidoreductase that catalyses the NADPH-dependent reduction of a wide variety of carbonyl compounds, including glucose. Its crystal structure has a single domain folded into an eight-stranded parallel α/β -barrel motif, with the substrate-binding site located in a cleft at the carboxy-terminal end of the β -barrel¹². Aldose

reductase has a low affinity (high K_m) for glucose, and at the normal glucose concentrations found in non-diabetics, metabolism of glucose by this pathway is a very small percentage of total glucose use. But in a hyperglycaemic environment, increased intracellular glucose results in its increased enzymatic conversion to the polyalcohol sorbitol, with concomitant decreases in NADPH. In the polyol pathway, sorbitol is oxidized to fructose by the enzyme sorbitol dehydrogenase, with NAD^+ reduced to NADH. Flux through this pathway during hyperglycaemia varies from 33% of total glucose use in the rabbit lens to 11% in human erythrocytes. Thus, the contribution of this pathway to diabetic complications may be very much species, site and tissue dependent (Fig. 1).

A number of mechanisms have been proposed to explain the potential detrimental effects of hyperglycaemia-induced increases in polyol pathway flux. These include sorbitol-induced osmotic stress, decreased $(Na^+ + K^+)ATPase$ activity, an increase in cytosolic NADH/ NAD^+ and a decrease in cytosolic NADPH. Sorbitol does not diffuse easily across cell membranes, and it was originally suggested that this resulted in osmotic damage to microvascular cells. Sorbitol concentrations measured in diabetic vessels and nerves are, however, far too low to cause osmotic damage.

Another early suggestion was that increased flux through the polyol pathway decreased $(Na^+ + K^+)ATPase$ activity. Although this decrease was originally thought to be mediated by polyol pathway-linked decreases in phosphatidylinositol synthesis, it has recently been shown to result from activation of PKC (see below). Hyperglycaemia-induced activation of PKC increases cytosolic phospholipase A_2 activity, which increases the production of two inhibitors of $(Na^+ + K^+)ATPase$ — arachidonate and PGE_2 (ref. 13).

More recently, it has been proposed that oxidation of sorbitol by NAD^+ increases the cytosolic NADH: NAD^+ ratio, thereby inhibiting activity of the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and increasing concentrations of triose phosphate¹⁴. Raised triose phosphate concentrations could increase formation of both methylglyoxal, a precursor of AGEs, and diacylglycerol (DAG) (through α -glycerol-3-phosphate), thus activating PKC (as discussed later). Although hyperglycaemia does increase the NADH: NAD^+ ratio in endothelial cells, this reflects a marked decrease in the absolute concentration of NAD^+ as a result of consumption by activated poly(ADP-ribose) polymerase (PARP), rather than reduction of NAD^+ to NADH¹⁵. Activation of PARP by hyperglycaemia is mediated by increased production of reactive oxygen species (T. Matsumura

et al., unpublished results). The source of hyperglycaemia-induced reactive oxygen species is discussed later.

It has also been proposed that reduction of glucose to sorbitol by NADPH consumes NADPH. As NADPH is required for regenerating reduced glutathione (GSH), this could induce or exacerbate intracellular oxidative stress (Fig. 1). Decreased levels of GSH have in fact been found in the lenses of transgenic mice that overexpress aldose reductase, and this is the most likely mechanism by which increased flux through the polyol pathway has deleterious consequences¹⁶. This conclusion is further supported by recent experiments with homozygous knockout mice deficient in aldose reductase, which showed that, in contrast to wild-type mice, diabetes neither decreased the GSH content of sciatic nerve nor reduced motor nerve conduction velocity (S. K. Chung, personal communication).

Studies of inhibition of the polyol pathway *in vivo* have yielded inconsistent results. In a five-year study in dogs, aldose reductase inhibition prevented diabetic neuropathy, but failed to prevent retinopathy or thickening of the capillary basement membrane in the retina, kidney and muscle¹⁷. Several negative clinical trials have questioned the relevance of this mechanism in humans¹⁸. The positive effect of aldose reductase inhibition on diabetic neuropathy has, however, been confirmed in humans in a rigorous multi-dose, placebo-controlled trial with the potent aldose reductase inhibitor zenarestat¹⁹.

Increased intracellular formation of advanced glycation end-products

AGEs are found in increased amounts in diabetic retinal vessels²⁰ and renal glomeruli²¹. They were originally thought to arise from non-enzymatic reactions between extracellular proteins and glucose. But the rate of AGE formation from glucose is orders of magnitude slower than the rate of AGE formation from glucose-derived dicarbonyl precursors generated intracellularly, and it now seems likely that intracellular hyperglycaemia is the primary initiating event in the formation of both intracellular and extracellular AGEs²². AGEs can arise from intracellular auto-oxidation of glucose to glyoxal²³, decomposition of the Amadori product (glucose-derived 1-amino-1-deoxyfructose lysine adducts) to 3-deoxyglucosone (perhaps accelerated by an amadoriase), and fragmentation of glyceraldehyde-3-phosphate and dihydroxyacetone phosphate to methylglyoxal²⁴. These reactive intracellular dicarbonyls — glyoxal, methylglyoxal and 3-deoxyglucosone — react with amino groups of intracellular and extracellular proteins to form AGEs. Methylglyoxal and glyoxal are detoxified by the glyoxalase system²⁴. All three AGE precursors are also substrates for other reductases²⁵.

The potential importance of AGEs in the pathogenesis of diabetic complications is indicated by the observation in animal models that two structurally unrelated AGE inhibitors partially prevented various functional and structural manifestations of diabetic microvascular disease in retina, kidney and nerve^{26–28}. In a large randomized, double-blind, placebo-controlled, multi-centre trial in type 1 diabetic patients with overt nephropathy, the AGE inhibitor aminoguanidine lowered total urinary protein and slowed progression of nephropathy, over and above the effects of existing optimal care. In addition, aminoguanidine reduced the progression of diabetic retinopathy (K. K. Bolton *et al.*, submitted).

Production of intracellular AGE precursors damages target cells by three general mechanisms (Fig. 2). First, intracellular proteins modified by AGEs have altered function. Second, extracellular matrix components modified by AGE precursors interact abnormally with other matrix components and with the receptors for matrix proteins (integrins) on cells. Third, plasma proteins modified by AGE precursors bind to AGE receptors on endothelial cells, mesangial cells and macrophages, inducing receptor-mediated production of reactive oxygen species. This AGE receptor ligation activates the pleiotropic transcription factor NF- κ B, causing pathological changes in gene expression.

In endothelial cells exposed to high glucose, intracellular AGE formation occurs within a week. Basic fibroblast growth factor is one

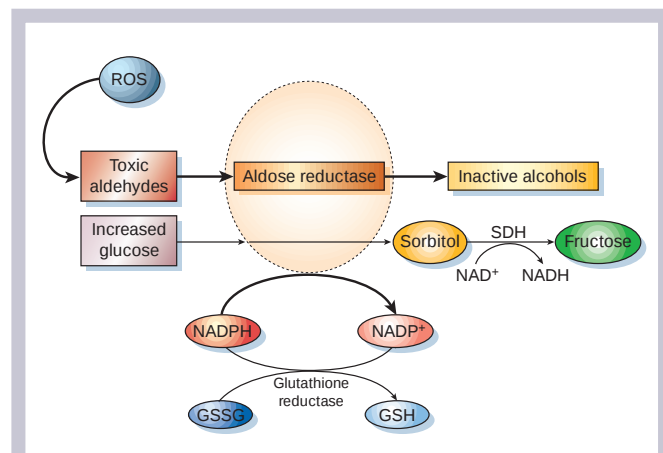
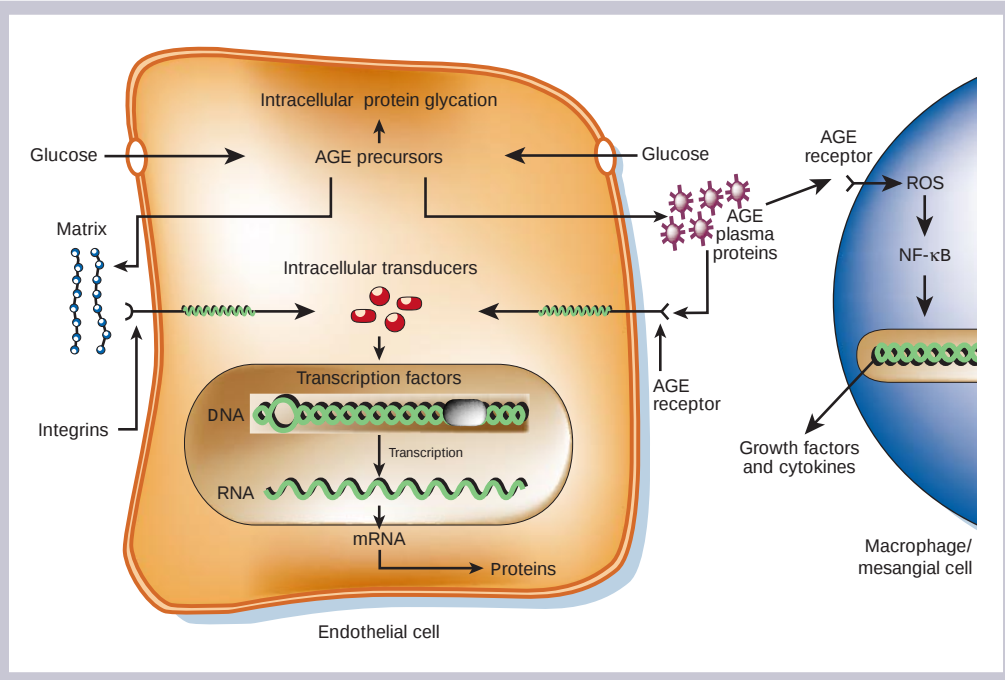


Figure 1 Aldose reductase and the polyol pathway. Aldose reductase reduces aldehydes generated by reactive oxygen species (ROS) to inactive alcohols, and glucose to sorbitol, using NADPH as a co-factor. In cells where aldose reductase activity is sufficient to deplete reduced glutathione (GSH), oxidative stress is augmented. Sorbitol dehydrogenase (SDH) oxidizes sorbitol to fructose using NAD^+ as a co-factor.

Figure 2 Mechanisms by which intracellular production of advanced glycation end-product (AGE) precursors damages vascular cells. Covalent modification of intracellular proteins by dicarbonyl AGE precursors alters several cellular functions. Modification of extracellular matrix proteins causes abnormal interactions with other matrix proteins and with integrins. Modification of plasma proteins by AGE precursors creates ligands that bind to AGE receptors, inducing changes in gene expression in endothelial cells, mesangial cells and macrophages.



of the main AGE-modified proteins in endothelial cells²⁹. Proteins involved in macromolecular endocytosis are also modified by AGEs, as the increase in endocytosis induced by hyperglycaemia is prevented by overexpression of the methylglyoxal-detoxifying enzyme glyoxalase I (ref. 30). Overexpression of glyoxalase I also completely prevents the hyperglycaemia-induced increase in expression of angiopoietin-2 in Muller cells (T. Matsumura *et al.*, unpublished results), a factor that has been implicated in both pericyte loss and capillary regression³¹.

AGE formation alters the functional properties of several important matrix molecules. On type I collagen, intermolecular crosslinking by AGEs induces an expansion of the molecular packing³². These AGE-induced crosslinks alter the function of intact vessels. For example, AGEs decrease elasticity in large vessels from diabetic rats, even after vascular tone is abolished, and increase fluid filtration across the carotid artery³³. AGE formation on type IV collagen from basement membrane inhibits lateral association of these molecules into a normal network-like structure by interfering with binding of the non-collagenous NC1 domain to the helix-rich domain³⁴. AGE formation on laminin causes decreased polymer self-assembly, decreased binding to type IV collagen, and decreased binding to heparan sulphate proteoglycan³⁵.

AGE formation on extracellular matrix not only interferes with matrix–matrix interactions, but also interferes with matrix–cell interactions. For example, AGE modification of type IV collagen's cell-binding domains decreases endothelial cell adhesion³⁶, and AGE modification of a growth-promoting sequence of six amino acids in the A chain of the laminin molecule markedly reduces neurite outgrowth³⁷.

Several cell-associated binding proteins for AGEs have been identified, including OST-48, 80K-H, galectin-3, the macrophage scavenger receptor type II and RAGE^{38–41}. Some of these are likely to contribute to clearance of AGEs, whereas others may underlie the sustained cellular perturbations mediated by binding of the AGE ligands. In cell-culture systems, the AGE receptors identified seem to mediate long-term effects of AGEs on key cellular targets of diabetic complications such as macrophages, glomerular mesangial cells and vascular endothelial cells, although not all these receptors bind proteins with physiological AGE-modification levels. These effects include expression of cytokines and growth factors by macrophages and mesangial cells (interleukin-1, insulin-like growth factor-I,

tumour necrosis factor- α , TGF- β , macrophage colony-stimulating factor, granulocyte–macrophage colony-stimulating factor and platelet-derived growth factor), and expression of pro-coagulatory and pro-inflammatory molecules by endothelial cells (thrombomodulin, tissue factor and the cell adhesion molecule VCAM-1)^{42–47}. In addition, binding of ligands to endothelial AGE receptors seems to mediate in part the hyperpermeability of the capillary wall induced by diabetes, probably through the induction of VEGF⁴⁸.

Consistent with this concept, blockade of one such receptor, RAGE, a member of the immunoglobulin superfamily with three immunoglobulin-like regions on a single polypeptide chain, suppressed macrovascular disease in an atherosclerosis-prone type 1 diabetic mouse model in a glucose- and lipid-independent fashion⁴⁹. Blockade of RAGE has also been shown to inhibit the development of diabetic nephropathy and periodontal disease, and to enhance wound repair in murine models. RAGE has been shown to mediate signal transduction, through generation of reactive oxygen species, which activates both the transcription factor NF- κ B, and p21^{Ras} (refs 50, 51). AGE signalling is blocked in cells by expression of RAGE antisense cDNA⁵² or an anti-RAGE ribozyme⁵³.

Activation of protein kinase C

The PKC family comprises at least eleven isoforms, nine of which are activated by the lipid second messenger DAG. Intracellular hyperglycaemia increases the amount of DAG in cultured microvascular cells and in the retina and renal glomeruli of diabetic animals. It seems to achieve this primarily by increasing *de novo* DAG synthesis from the glycolytic intermediate dihydroxyacetone phosphate, through reduction of the latter to glycerol-3-phosphate and stepwise acylation⁵⁴. Increased *de novo* synthesis of DAG activates PKC both in cultured vascular cells⁵⁵ and in retina and glomeruli of diabetic animals⁵⁴. The β - and δ -isoforms of PKC are activated primarily, but increases in other isoforms have also been found, such as PKC- α and - ϵ isoforms in the retina⁵⁴ and PKC- α and - β in glomeruli⁵⁶ of diabetic rats. Hyperglycaemia may also activate PKC isoforms indirectly through both ligation of AGE receptors⁵⁷ and increased activity of the polyol pathway⁵⁸, presumably by increasing reactive oxygen species.

In early experimental diabetes, activation of PKC- β isoforms has been shown to mediate retinal and renal blood flow abnormalities⁵⁹, perhaps by depressing nitric oxide production and/or increasing endothelin-1 activity (Fig. 3). Abnormal activation of PKC has been

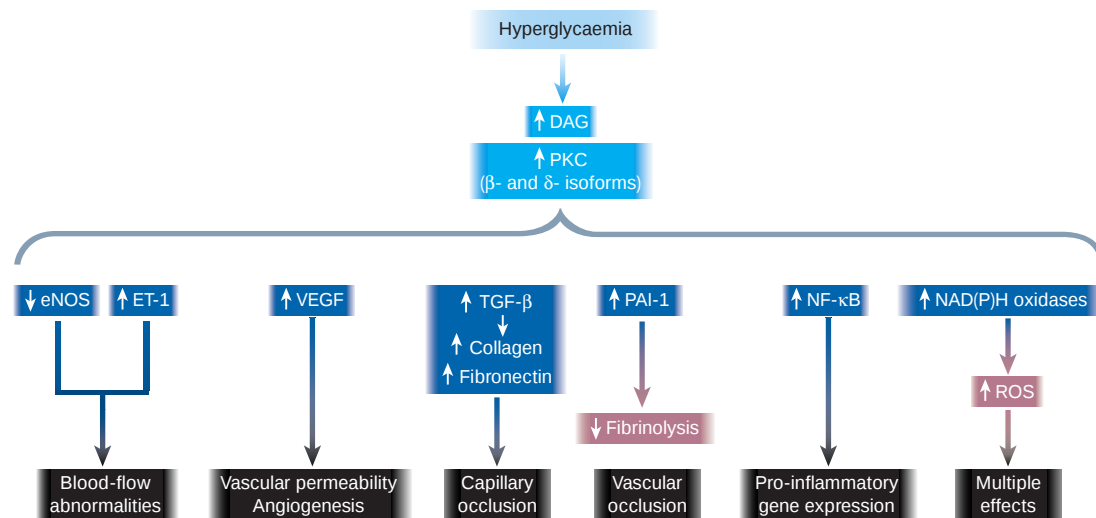


Figure 3 Consequences of hyperglycaemia-induced activation of protein kinase C (PKC). Hyperglycaemia increases diacylglycerol (DAG) content, which activates PKC, primarily the β - and δ -isoforms. Activation of PKC has a number of pathogenic consequences by affecting expression of endothelial nitric oxide synthetase (eNOS),

endothelin-1 (ET-1), vascular endothelial growth factor (VEGF), transforming growth factor- β (TGF- β) and plasminogen activator inhibitor-1 (PAI-1), and by activating NF- κ B and NAD(P)H oxidases.

implicated in the decreased glomerular production of nitric oxide induced by experimental diabetes⁶⁰, and in the decreased production of nitric oxide in smooth muscle cells that is induced by hyperglycaemia⁶¹. Activation of PKC also inhibits insulin-stimulated expression of the messenger RNA for endothelial nitric oxide synthase (eNOS) in cultured endothelial cells⁶². Hyperglycaemia increases endothelin-1-stimulated MAP-kinase activity in glomerular mesangial cells by activating PKC isoforms⁶³. The increased

permeability of endothelial cells induced by high glucose in cultured cells is mediated by activation of PKC- α , however⁶⁴. Activation of PKC by raised glucose also induces expression of the permeability-enhancing factor VEGF in smooth muscle cells⁶⁵.

In addition to affecting hyperglycaemia-induced abnormalities of blood flow and permeability, activation of PKC contributes to increased microvascular matrix protein accumulation by inducing expression of TGF- β 1, fibronectin and type IV collagen both in cultured mesangial cells⁶⁶ and in glomeruli of diabetic rats⁶⁷. This effect seems to be mediated through inhibition of nitric oxide production by PKC⁶⁸. But hyperglycaemia-induced expression of laminin C1 in cultured mesangial cells is independent of PKC activation⁶⁹. Hyperglycaemia-induced activation of PKC has also been implicated in the overexpression of the fibrinolytic inhibitor PAI-1 (ref. 70), the activation of NF- κ B in cultured endothelial cells and vascular smooth muscle cells^{71,72}, and in the regulation and activation of various membrane-associated NAD(P)H-dependent oxidases.

Treatment with an inhibitor specific for PKC- β significantly reduced PKC activity in the retina and renal glomeruli of diabetic animals. Concomitantly, treatment significantly reduced diabetes-induced increases in retinal mean circulation time, normalized increases in glomerular filtration rate and partially corrected urinary albumin excretion. Treatment of a mouse model of type 2 diabetes (*db/db*) with a β -isoform-specific PKC inhibitor ameliorated accelerated glomerular mesangial expansion⁷³.

Increased flux through the hexosamine pathway

Shunting of excess intracellular glucose into the hexosamine pathway might also cause several manifestations of diabetic complications⁷⁴. In this pathway, fructose-6-phosphate is diverted from glycolysis to provide substrates for reactions that require UDP-*N*-acetylglucosamine, such as proteoglycan synthesis and the formation of *O*-linked glycoproteins (Fig. 4). Inhibition of the rate-limiting enzyme in the conversion of glucose to glucosamine — glutamine:fructose-6-phosphate amidotransferase (GFAT) — blocks hyperglycaemia-induced increases in the transcription of TGF- α , TGF- β 1 (ref. 74) and PAI-1 (ref. 10). This pathway is also important role in hyperglycaemia-induced and fat-induced insulin resistance^{75,76}.

The mechanism by which increased flux through the hexosamine pathway mediates hyperglycaemia-induced increases in gene

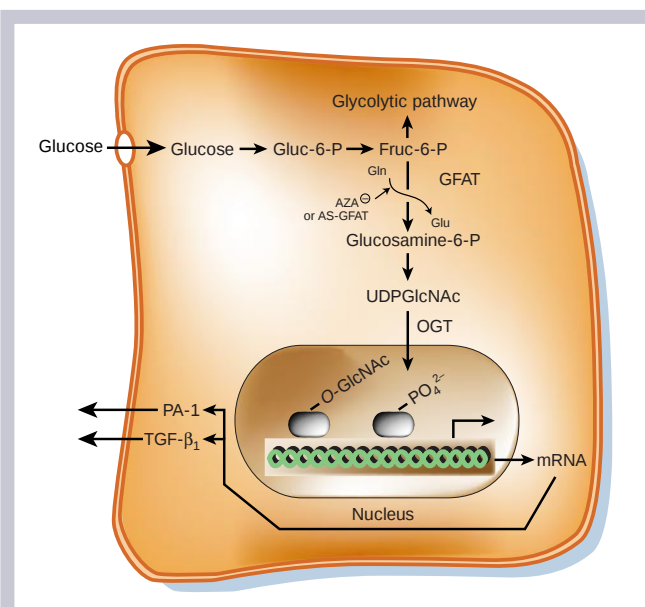


Figure 4 The hexosamine pathway. The glycolytic intermediate fructose-6-phosphate (Fru-6-P) is converted to glucosamine-6-phosphate by the enzyme glutamine:fructose-6-phosphate amidotransferase (GFAT). Intracellular glucosylation by the addition of *N*-acetylglucosamine (GlcNAc) to serine and threonine is catalysed by the enzyme *O*-GlcNAc transferase (OGT). Increased donation of GlcNAc moieties to serine and threonine residues of transcription factors such as Sp1, often at phosphorylation sites, increases the production of factors as PAI-1 and TGF- β 1. AZA, azaserine; AS-GFAT, antisense to GFAT.

transcription is not certain, but the observation that binding sites for the transcription factor Sp1 regulate hyperglycaemia-induced activation of the PAI-1 promoter in vascular smooth muscle cells⁷⁷ suggested that covalent modification of Sp1 by *N*-acetylglucosamine (GlcNAc) might explain the link between activation of the hexosamine pathway and hyperglycaemia-induced changes in transcription of the gene for PAI-1. Glucosamine itself was subsequently shown to activate the PAI-1 promoter through Sp1 sites in glomerular mesangial cells⁷⁸. The glycosylated form of Sp1 seems to be more transcriptionally active than the deglycosylated form⁷⁹. A fourfold increase in *O*-acetylglucosaminylation of Sp1 caused by inhibition of the enzyme *O*-GlcNAc- β -*N*-acetylglucosaminidase resulted in a reciprocal 30% decrease in the level of serine-threonine phosphorylation of Sp1, supporting the concept that *O*-acetylglucosaminylation and phosphorylation compete for the same sites on this protein⁸⁰.

Recently, hyperglycaemia was shown to induce a 2.4-fold increase in hexosamine pathway activity in aortic endothelial cells, resulting in a 1.7-fold increase in Sp1 *O*-linked GlcNAc and a 70–80% decrease in Sp1 *O*-linked phosphothreonine and phosphoserine¹⁰. Concomitantly, hyperglycaemia resulted in a 3.8-fold increase in expression from an 85-base-pair truncated PAI-1 promoter-luciferase reporter DNA containing two Sp1 sites, but failed to increase expression when the two Sp1 sites were mutated¹⁰.

Modification of Sp1 by GlcNAc may regulate other glucose-responsive genes in addition to that for PAI-1. As virtually every RNA polymerase II transcription factor examined has been found to be *O*-acetylglucosaminylated⁸¹, it is possible that reciprocal modification by *O*-acetylglucosaminylation and phosphorylation of transcription factors other than Sp1 may function as a more generalized mechanism for regulating glucose-responsive gene transcription.

In addition to transcription factors, many other nuclear and cytoplasmic proteins are dynamically modified by *O*-linked GlcNAc, and may show reciprocal modification by phosphorylation in a manner analogous to Sp1 (ref. 81). One example relevant to diabetic complications is the inhibition of eNOS activity by hyperglycaemia-induced *O*-acetylglucosaminylation at the Akt site of the eNOS protein⁸². Other examples might be various PKC isoforms, which are activated by glucosamine without membrane translocation (H. J. Goldberg, C. J. Whiteside & G. Fantus, personal communication).

Thus, activation of the hexosamine pathway by hyperglycaemia may result in many changes in both gene expression and protein function, which together contribute to the pathogenesis of diabetic complications.

A common element linking hyperglycaemia-induced damage

Although specific inhibitors of aldose reductase activity, AGE formation, PKC activation and the hexosamine pathway each ameliorate various diabetes-induced abnormalities in cell culture and animal models, there has been no apparent common element linking the four mechanisms of hyperglycaemia-induced damage^{17,26–28,59,83}. This issue has now been resolved by the recent discovery that each of the four different pathogenic mechanisms reflects a single hyperglycaemia-induced process: overproduction of superoxide by the mitochondrial electron-transport chain^{10,84}. Many studies have shown that diabetes and hyperglycaemia increase oxidative stress⁸⁵, but neither the underlying mechanism nor the consequences for other pathways of hyperglycaemic damage were known.

To understand how hyperglycaemia increases the production of reactive oxygen species inside cultured aortic endothelial cells⁸⁶, a brief overview of glucose metabolism is helpful (Box 1). When the electrochemical potential difference generated by the proton gradient across the inner mitochondrial membrane is high, the lifetime of superoxide-generating electron-transport intermediates such as ubiquinone is prolonged (Fig. 5). There seems to be a threshold value above which superoxide production is markedly increased⁸⁷. Du *et al.*⁸² have found that hyperglycaemia increases the

Box 1

Overview of glucose metabolism

Intracellular glucose oxidation begins with glycolysis in the cytoplasm, which generates NADH and pyruvate. Cytoplasmic NADH can donate reducing equivalents to the mitochondrial electron-transport chain by way of two shuttle systems, or it can reduce pyruvate to lactate, which exits the cell to provide substrate for hepatic gluconeogenesis. Pyruvate can also be transported into the mitochondria, where it is oxidized by the tricarboxylic acid (TCA) cycle to produce CO₂, H₂O, four molecules of NADH and one molecule of FADH₂. Mitochondrial NADH and FADH₂ provide energy for ATP production through oxidative phosphorylation by the electron-transport chain.

Electron flow through the mitochondrial electron-transport chain (Fig. 5) is carried out by four inner membrane-associated enzyme complexes, plus cytochrome *c* and the mobile electron carrier ubiquinone. NADH derived from both cytosolic glucose oxidation and mitochondrial TCA cycle activity donates electrons to NADH:ubiquinone oxidoreductase (complex I). Complex I ultimately transfers its electrons to ubiquinone. Ubiquinone can also be reduced by electrons donated from several FADH₂-containing dehydrogenases, including succinate:ubiquinone oxidoreductase (complex II) and glycerol-3-phosphate dehydrogenase. Electrons from reduced ubiquinone are then transferred to ubiquinol:cytochrome *c* oxidoreductase (complex III) by the ubiquinol:cytochrome *c* radical-generating Q cycle. Electron transport then proceeds through cytochrome *c*, cytochrome *c* oxidase (complex IV) and, finally, molecular oxygen (O₂). Electron transfer through complexes I, III and IV generates a proton gradient that drives ATP synthase (complex V).

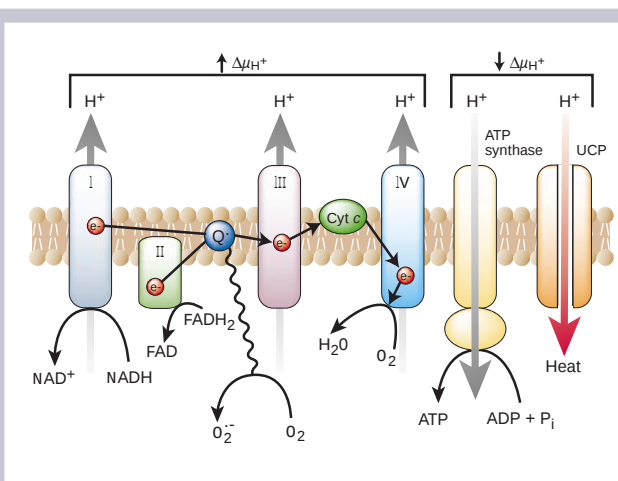


Figure 5 Production of superoxide by the mitochondrial electron-transport chain. Increased hyperglycaemia-derived electron donors from the TCA cycle (NADH and FADH₂) generate a high mitochondrial membrane potential ($\Delta\mu_{H^+}$) by pumping protons across the mitochondrial inner membrane. This inhibits electron transport at complex III, increasing the half-life of free-radical intermediates of coenzyme Q (ubiquinone), which reduce O₂ to superoxide.

proton gradient above this threshold value as a result of overproduction of electron donors by the TCA cycle. This, in turn, causes a marked increase in the production of superoxide by endothelial cells. Overexpression of manganese superoxide dismutase (MnSOD), the mitochondrial form of superoxide dismutase, abolished the signal generated by reactive oxygen species, and overexpression of uncoupling protein-1 (UCP-1) collapsed the proton electrochemical

gradient and prevented hyperglycaemia-induced overproduction of reactive oxygen species.

Inhibition by MnSOD or UCP-1 of hyperglycaemia-induced overproduction of mitochondrial superoxide completely prevented an increase in polyol pathway flux, increased intracellular AGE formation, increased PKC activation and an increase in hexosamine pathway activity in endothelial cells.

As hyperglycaemia-induced overproduction of mitochondrial superoxide induces a 66% decrease in GAPDH activity¹⁰, the effect of hyperglycaemia on polyol pathway flux may reflect the accumulation of glycolytic metabolites, including glucose, upstream of GAPDH (Fig. 6). Although the reversible inhibition of GAPDH by reactive oxygen species has been well described, the inhibition of GAPDH by hyperglycaemia-induced reactive oxygen species may reflect the resulting activation of PARP and depletion of NAD⁺ (ref. 15), rather than direct oxidative inactivation of the enzyme, as intracellular GSH levels remain high.

In regard to AGEs, methylglyoxal-derived AGE, the primary intracellular AGE induced by hyperglycaemia³⁰, is formed by fragmentation of triose phosphates. Thus the effect of hyperglycaemia on intracellular AGE formation also probably reflects increased triose phosphate levels resulting from inhibition of GAPDH by mitochondrial overproduction of reactive oxygen species (Fig. 6)¹⁰. This hypothesis is supported by the observation that GAPDH antisense oligonucleotides caused identical intracellular increases in AGE in the absence of hyperglycaemia (M.B. *et al.*, unpublished results).

Hyperglycaemia activates PKC by increasing the *de novo* synthesis of DAG⁵⁴, so the effect of hyperglycaemia on PKC activation probably reflects increased dihydroxyacetone phosphate levels resulting from inhibition of GAPDH by reactive oxygen species (Fig. 6)¹⁰. That GAPDH antisense oligonucleotides also caused activation of PKC in the presence of physiological glucose concentrations (M. B. *et al.*, unpublished results) supports this hypothesis.

Because hyperglycaemia increases hexosamine pathway flux by providing more fructose-6-phosphate for GFAT — the rate-limiting enzyme of the pathway — the effect of hyperglycaemia on hexosamine pathway flux probably reflects increased fructose-6-phosphate levels, resulting from inhibition of GAPDH by reactive oxygen species (Fig. 6)¹⁰. GAPDH antisense oligonucleotides also caused identical increases in hexosamine pathway flux in the absence of hyperglycaemia (M. B. *et al.*, unpublished results). Hyperglycaemia-induced activation of the redox-sensitive pleiotropic transcription factor NF- κ B was also prevented by inhibition of mitochondrial superoxide overproduction⁸⁴.

Overexpression of UCP-1 or MnSOD corrects a variety of hyperglycaemia-induced phenotypes in target cells of diabetic complications. In cultured glomerular mesangial cells, overexpression of MnSOD suppresses the increase in collagen synthesis induced by high glucose⁸⁸. In dorsal root ganglion (DRG) neurons from both wild-type and MnSOD^{+/-} mice, overexpression of MnSOD decreases hyperglycaemia-induced programmed cell death, and in embryonic rat DRG neurons, overexpression of UCP-1 inhibits cleavage of programmed cell death effector caspases (J. W. Russell, personal communication). In aortic endothelial cells, overexpression of either UCP-1 or MnSOD completely blocks hyperglycaemia-induced monocyte adhesion (J. L. Nadler & C. C. Hedrick, personal communication), prevents hyperglycaemia-induced inhibition of the anti-atherogenic enzyme prostacyclin synthetase, and prevents hyperglycaemia-induced inhibition of peroxisome proliferator-activated receptor- γ activation (M. B. *et al.*, unpublished data). Overexpression of either MnSOD or UCP-1 also prevents inhibition of eNOS activity by hyperglycaemia⁸². In platelets, chemical uncouplers or SOD mimetics both prevent potentiation by hyperglycaemia of collagen-induced platelet activation and aggregation⁸⁹.

In transgenic mice overexpressing human cytoplasmic Cu²⁺/Zn²⁺ SOD, and in which diabetes was induced by streptozotocin (STZ) treatment, albuminuria, glomerular hypertrophy and glomerular

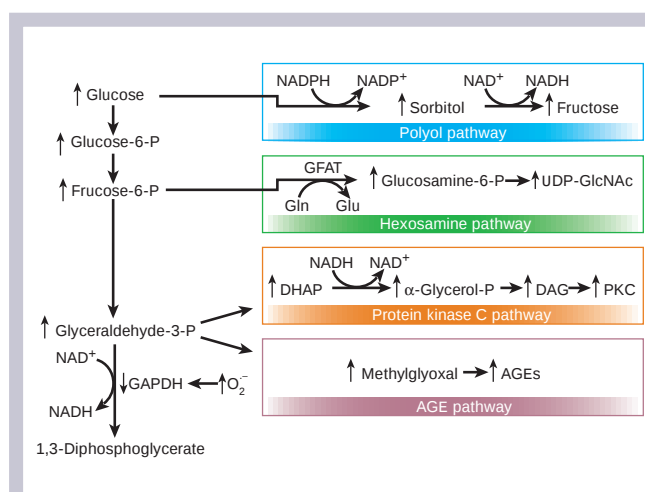


Figure 6 Potential mechanism by which hyperglycaemia-induced mitochondrial superoxide overproduction activates four pathways of hyperglycaemic damage. Excess superoxide partially inhibits the glycolytic enzyme GAPDH, thereby diverting upstream metabolites from glycolysis into pathways of glucose overutilization. This results in increased flux of dihydroxyacetone phosphate (DHAP) to DAG, an activator of PKC, and of triose phosphates to methylglyoxal, the main intracellular AGE precursor. Increased flux of fructose-6-phosphate to UDP-*N*-acetylglucosamine increases modification of proteins by *O*-linked *N*-acetylglucosamine (GlcNAc) and increased glucose flux through the polyol pathway consumes NADPH and depletes GSH.

content of TGF- β and collagen type IV were all attenuated compared to wild-type littermates after 4 months of diabetes. The wild-type STZ-diabetic mice developed modest increases in fractional mesangial volume after 8 months of diabetes, and this change was also suppressed in the SOD transgenic diabetic mice⁹⁰. Overexpression of the human SOD1 transgene in *db/db* diabetic mice similarly attenuated the extensive expansion of the glomerular mesangial matrix that was otherwise evident by age 5 months in the non-transgenic *db/db* littermates (F. R. DeRubertis, P. A. Craven, M. F. Melhem, H. Liachenko & S. L. Phillips, unpublished observations).

Future directions

The discovery that each of the four main mechanisms implicated in the pathogenesis of diabetic complications reflects a single hyperglycaemia-induced process provides a new conceptual framework for future research, although clinical trials will be necessary to show that the results from cell culture and animal studies are applicable to humans. Three general areas are of great importance to a more complete understanding of the molecular and cell biology of diabetic complications.

First is the phenomenon of so-called hyperglycaemic memory. This refers to the persistence or progression of hyperglycaemia-induced microvascular alterations during subsequent periods of normal glucose homeostasis. The most striking example of this occurred in the eyes of diabetic dogs during a post-hyperglycaemic period of euglycaemia⁹¹. After 2.5 years of elevated glucose, the eyes were histologically normal. But after a subsequent 2.5-year period of normal glycaemia, the eyes developed severe retinopathy. Results from the Epidemiology of Diabetes Interventions and Complications study indicate that hyperglycaemic memory also occurs in human patients. The effects of intensive and conventional therapy on the occurrence and severity of post-study retinopathy and nephropathy were shown to persist for four years after the Diabetes Control and Complications Trial (DCCT), despite nearly identical glycosylated haemoglobin values during the 4-year follow-up⁹². Hyperglycaemia-induced mitochondrial superoxide production may provide an explanation for the development of complications

during post-hyperglycaemic periods of normal glycaemia. Hyperglycaemia-induced increases in superoxide would not only increase polyol pathway flux, AGE formation, PKC activity and hexosamine pathway flux, but might also induce mutations in mitochondrial DNA. Defective subunits of the electron-transport complexes encoded by mutated mitochondrial DNA could eventually cause increased superoxide production at physiological concentrations of glucose, with resulting continued activation of the four pathways despite the absence of hyperglycaemia.

The second general area concerns the genetic determinants of susceptibility to both microvascular and macrovascular complications. Their role in microvascular complications is supported by familial clustering of diabetic nephropathy and retinopathy. In two studies of families in which two or more siblings had type 1 diabetes, if one diabetic sibling had advanced diabetic nephropathy, the other diabetic sibling had a nephropathy risk of 83% or 72%, whereas the risk was only 17% or 22% if the index case did not have diabetic nephropathy⁹³. The DCCT reported familial clustering of retinopathy with an odds ratio of 5.4 for the risk of severe retinopathy in diabetic relatives of retinopathy-positive subjects from the conventional treatment group compared with subjects with no retinopathy⁹⁴. For macrovascular complications, coronary artery calcification (an indicator of subclinical atherosclerosis) also shows familial clustering, with an estimated heritability of at least 40% (ref. 95). Thus, gene-mapping studies designed to identify genes that predispose to complications, as well as the interaction of these genes with metabolic factors, are warranted.

Finally, the paradigm discussed in this review suggests that interrupting the overproduction of superoxide by the mitochondrial electron-transport chain would normalize polyol pathway flux, AGE formation, PKC activation, hexosamine pathway flux and NF- κ B activation. But it might be difficult to accomplish this using conventional antioxidants, as these scavenge reactive oxygen species in a stoichiometric manner. Thus, although long-term administration of a multi-antioxidant diet inhibited the development of early diabetic retinopathy in rats⁹⁶, and vitamin C improved endothelium-dependent vasodilation in diabetic patients⁹⁷, low-dose vitamin E failed to alter the risk of cardiovascular and renal disease in patients with diabetes⁹⁸. New, low-molecular-mass compounds that act as SOD or catalase mimetics have the theoretical advantage of scavenging reactive oxygen species continuously by acting as catalysts with efficiencies approaching those of the native enzymes⁹⁹. Such compounds normalize diabetes-induced inhibition of aortic prostacyclin synthetase in animals (M. B. *et al.*, unpublished results), and significantly improve diabetes-induced decreases in endoneurial blood flow and motor nerve conduction velocity¹⁰⁰. These and other agents discovered using high-throughput chemical and biological methods might have unique clinical efficacy in preventing the development and progression of diabetic complications. □

1. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.* **329**, 977–986 (1993).
2. UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* **352**, 837–853 (1998).
3. Wei, M., Gaskill, S. P., Haffner, S. M. & Stern, M. P. Effects of diabetes and level of glycemia on all-cause and cardiovascular mortality. The San Antonio Heart Study. *Diabetes Care* **7**, 1167–1172 (1998).
4. Ebara, T. *et al.* Delayed catabolism of apoB-48 lipoproteins due to decreased heparan sulfate proteoglycan production in diabetic mice. *J. Clin. Invest.* **105**, 1807–1818 (2000).
5. Ginsberg, H. N. Insulin resistance and cardiovascular disease. *J. Clin. Invest.* **106**, 453–458 (2000).
6. Lusis, A. J. Atherosclerosis. *Nature* **407**, 233–241 (2000).
7. Hsueh, W. A. & Law, R. E. Cardiovascular risk continuum: implications of insulin resistance and diabetes. *Am. J. Med.* **105**, 45–145 (1998).
8. Jiang, Z. Y. *et al.* Characterization of selective resistance to insulin signaling in the vasculature of obese Zucker (fa/fa) rats. *J. Clin. Invest.* **104**, 447–457 (1999).
9. Williams, S. B. *et al.* Acute hyperglycemia attenuates endothelium-dependent vasodilation in humans in vivo. *Circulation* **97**, 1695–1701 (1998).
10. Du, X. L. *et al.* Hyperglycemia-induced mitochondrial superoxide overproduction activates the hexosamine pathway and induces plasminogen activator inhibitor-1 expression by increasing Sp1 glycosylation. *Proc. Natl Acad. Sci. USA* **97**, 12222–12226 (2000).
11. Temelkova-Kurktschiev, T. S. *et al.* Postchallenge plasma glucose and glycemic spikes are more

- strongly associated with atherosclerosis than fasting glucose or HbA1c levels. *Diabetes Care* **12**, 1830–1834 (2000).
12. Wilson, D. K., Bohren, K. M., Gabbay, K. H. & Quiocho, F. A. An unlikely sugar substrate site in the 1.65 Å structure of the human aldose reductase holoenzyme implicated in diabetic complications. *Science* **257**, 81–84 (1992).
13. Xia, P., Kramer, R. M. & King, G. L. Identification of the mechanism for the inhibition of Na,K-adenosine triphosphatase by hyperglycemia involving activation of protein kinase C and cytosolic phospholipase A2. *J. Clin. Invest.* **96**, 733–740 (1995).
14. Williamson, J. R. *et al.* Hyperglycemic pseudohypoxia and diabetic complications. *Diabetes* **42**, 801–813 (1993).
15. Garcia Soriano, F. *et al.* Diabetic endothelial dysfunction: the role of poly(ADP-ribose) polymerase activation. *Nature Med.* **7**, 108–113 (2001).
16. Lee, A. Y. & Chung, S. S. Contributions of polyol pathway to oxidative stress in diabetic cataract. *FASEB J.* **13**, 23–30 (1999).
17. Engerman, R. L., Kern, T. S. & Larson, M. E. Nerve conduction and aldose reductase inhibition during 5 years of diabetes or galactosaemia in dogs. *Diabetologia* **37**, 141–144 (1994).
18. Sorbinil Retinopathy Trial Research Group. A randomized trial of sorbinil, an aldose reductase inhibitor, in diabetic retinopathy. *Arch. Ophthalmol.* **108**, 1234–1244 (1990).
19. Greene, D. A., Arezzo, J. C. & Brown, M. B. Effect of aldose reductase inhibition on nerve conduction and morphometry in diabetic neuropathy. Zenarestat Study Group. *Neurology* **53**, 580–591 (1999).
20. Stitt, A. W. *et al.* Advanced glycation end products (AGEs) co-localize with AGE receptors in the retinal vasculature of diabetic and of AGE-infused rats. *Am. J. Pathol.* **150**, 523–528 (1997).
21. Horie, K. *et al.* Immunohistochemical colocalization of glycoxidation products and lipid peroxidation products in diabetic renal glomerular lesions. Implication for glycoxidative stress in the pathogenesis of diabetic nephropathy. *J. Clin. Invest.* **100**, 2995–2999 (1997).
22. Degenhardt, T. P., Thorpe, S. R. & Baynes, J. W. Chemical modification of proteins by methylglyoxal. *Cell Mol. Biol.* **44**, 1139–1145 (1998).
23. Wells-Knecht, K. J. *et al.* Mechanism of autooxidative glycosylation: identification of glyoxal and arabinose as intermediates in the autooxidative modification of proteins by glucose. *Biochemistry* **34**, 3702–3709 (1995).
24. Thornalley, P. J. The glyoxalase system: new developments towards functional characterization of a metabolic pathway fundamental to biological life. *Biochem. J.* **269**, 1–11 (1990).
25. Suzuki, K. *et al.* Overexpression of aldehyde reductase protects PC12 cells from the cytotoxicity of methylglyoxal or 3-deoxyglucosone. *J. Biochem.* **123**, 353–357 (1998).
26. Soulis-Liparota, D., Cooper, M., Papazoglou, D., Clarke, B. & Jerums, G. Retardation by aminoguanidine of development of albuminuria, mesangial expansion, and tissue fluorescence in streptozocin-induced diabetic rat. *Diabetes* **40**, 1328–1334 (1991).
27. Nakamura, S. *et al.* Progression of nephropathy in spontaneous diabetic rats is prevented by OPB-9195, a novel inhibitor of advanced glycation. *Diabetes* **46**, 895–899 (1997).
28. Hammes, H.-P. *et al.* Aminoguanidine treatment inhibits the development of experimental diabetic retinopathy. *Proc. Natl Acad. Sci. USA* **88**, 11555–11559 (1991).
29. Giordano, I., Edelstein, D. & Brownlee, M. Nongenymatic glycosylation in vitro and in bovine endothelial cells alters basic fibroblast growth factor activity. A model for intracellular glycosylation in diabetes. *J. Clin. Invest.* **94**, 110–117 (1994).
30. Shinohara, M. *et al.* Overexpression of glyoxalase-I in bovine endothelial cells inhibits intracellular advanced glycation endproduct formation and prevents hyperglycemia-induced increases in macromolecular endocytosis. *J. Clin. Invest.* **101**, 1142–1147 (1998).
31. Maisonnepierre, P. C. *et al.* Angiopoietin-2, a natural antagonist for Tie2 that disrupts in vivo angiogenesis. *Science* **277**, 55–60 (1997).
32. Tanaka, S., Avigad, G., Brodsky, B. & Eikenberry, E. F. Glycation induces expansion of the molecular packing of collagen. *J. Mol. Biol.* **203**, 495–505 (1988).
33. Huijberts, M. S. P. *et al.* Aminoguanidine treatment increases elasticity and decreases fluid filtration of large arteries from diabetic rats. *J. Clin. Invest.* **92**, 1407–1411 (1993).
34. Tsilibary, E. C. *et al.* The effect of nongenymatic glycosylation the binding of the main noncollagenous NC1 domain to type IV collagen. *J. Biol. Chem.* **263**, 4302–4308 (1990).
35. Charonis, A. S. *et al.* Laminin alterations after in vitro nongenymatic glycosylation. *Diabetes* **39**, 807–814 (1988).
36. Haitoglou, C. S., Tsilibary, E. C., Brownlee, M. & Charonis, A. S. Altered cellular interactions between endothelial cells and nongenymatically glycosylated laminin/type IV collagen. *J. Biol. Chem.* **267**, 12404–12407 (1992).
37. Federoff, H. J., Lawrence, D. & Brownlee, M. Nongenymatic glycosylation of laminin and the laminin peptide C1KVAVS inhibits neurite outgrowth. *Diabetes* **42**, 509–513 (1993).
38. Li, Y. M. *et al.* Molecular identity and cellular distribution of advanced glycation endproduct receptors: relationship of p60 to OST-48 and p90 to 80K-H membrane proteins. *Proc. Natl Acad. Sci. USA* **93**, 11047–11052 (1996).
39. Smedsrod, B. *et al.* Advanced glycation end products are eliminated by scavenger-receptor-mediated endocytosis in hepatic sinusoidal kupffer and endothelial cells. *Biochem. J.* **322**, 567–573 (1997).
40. Vlassara, H. *et al.* Identification of galectin-3 as a high-affinity binding protein for advanced glycation end products (AGE): a new member of the AGE-receptor complex. *Mol. Med.* **1**, 634–646 (1995).
41. Neep, M. *et al.* Cloning and expression of RAGE: a cell surface receptor for advanced glycosylation end products of proteins. *J. Biol. Chem.* **267**, 14998–15004 (1992).
42. Vlassara, H. *et al.* Cachectin/TNF and IL-1 induced by glucose-modified proteins: role in normal tissue remodeling. *Science* **240**, 1546–1548 (1988).
43. Kirstein, M., Aston, C., Hintz, R. & Vlassara, H. Receptor-specific induction of insulin-like growth factor I in human monocytes by advanced glycosylation end product-modified proteins. *J. Clin. Invest.* **90**, 439–446 (1992).
44. Abordo, E. A., Westwood, M. E. & Thornalley, P. J. Synthesis and secretion of macrophage colony stimulating factor by mature human monocytes and human monocytic THP-1 cells induced by human serum albumin derivatives modified with methylglyoxal and glucose-derived advanced glycation endproducts. *Immunol. Lett.* **53**, 7–13 (1996).
45. Skolnik, E. Y. *et al.* Human and rat mesangial cell receptors for glucose-modified proteins: potential role in kidney tissue remodelling and diabetic nephropathy. *J. Exp. Med.* **174**, 931–939 (1991).
46. Doi, T. *et al.* Receptor specific increase in extracellular matrix productions in mouse mesangial cells by advanced glycosylation end products is mediated via platelet derived growth factor. *Proc. Natl Acad. Sci. USA* **89**, 2873–2877 (1992).

47. Schmidt, A. M. *et al.* Advanced glycation endproducts interacting with their endothelial receptor induce expression of vascular cell adhesion molecule-1 (VCAM-1) in cultured human endothelial cells and in mice: a potential mechanism for the accelerated vasculopathy of diabetes. *J. Clin. Invest.* **96**, 1395–1403 (1995).
48. Lu, M. *et al.* Advanced glycation end products increase retinal vascular endothelial growth factor expression. *J. Clin. Invest.* **101**, 1219–1224 (1998).
49. Park, L. *et al.* Suppression of accelerated diabetic atherosclerosis by the soluble receptor for advanced glycation endproducts. *Nature Med.* **4**, 1025–1031 (1998).
50. Yan, S. D. *et al.* Enhanced cellular oxidant stress by the interaction of advanced glycation end products with their receptors/ligands. *J. Biol. Chem.* **269**, 9889–9897 (1994).
51. Lander, H. M. *et al.* Activation of the receptor for advanced glycation end products triggers a p21(ras)-dependent mitogen-activated protein kinase pathway regulated by oxidant stress. *J. Biol. Chem.* **272**, 17810–17814 (1997).
52. Yamagishi, S. *et al.* Advanced glycation endproducts inhibit prostacyclin production and induce plasminogen activator inhibitor-1 in human microvascular endothelial cells. *Diabetologia* **41**, 1435–1441 (1998).
53. Tsuji, H. *et al.* Ribozyme targeting of receptor for advanced glycation end products in mouse mesangial cells. *Biochem. Biophys. Res. Commun.* **245**, 583–588 (1998).
54. Koya, D. & King, G. L. Protein kinase C activation and the development of diabetic complications. *Diabetes* **47**, 859–866 (1998).
55. Xia, P. *et al.* Characterization of the mechanism for the chronic activation of diacylglycerol-protein kinase C pathway in diabetes and hypergalactosemia. *Diabetes* **43**, 1122–1129 (1994).
56. Koya, D. *et al.* Characterization of protein kinase C beta isoform activation on the gene expression of transforming growth factor-beta, extracellular matrix components, and prostanoids in the glomeruli of diabetic rats. *J. Clin. Invest.* **100**, 115–126 (1997).
57. Portilla, D. *et al.* Etomoxir -induced PPARalpha-modulated enzymes protect during acute renal failure. *Am. J. Physiol. Renal Physiol.* **278**, F667–F675 (2000).
58. Keogh, R. J., Dunlop, M. E. & Larkins, R. G. Effect of inhibition of aldose reductase on glucose flux, diacylglycerol formation, protein kinase C, and phospholipase A2 activation. *Metabolism* **46**, 41–47 (1997).
59. Ishii, H. *et al.* Amelioration of vascular dysfunctions in diabetic rats by an oral PKC beta inhibitor. *Science* **272**, 728–731 (1996).
60. Craven, P. A., Studer, R. K. & DeRubertis, F. R. Impaired nitric oxide-dependent cyclic guanosine monophosphate generation in glomeruli from diabetic rats. Evidence for protein kinase C-mediated suppression of the cholinergic response. *J. Clin. Invest.* **93**, 311–320 (1994).
61. Ganz, M. B. & Seftel, A. Glucose-induced changes in protein kinase C and nitric oxide are prevented by vitamin E. *Am. J. Physiol.* **278**, E146–E152 (2000).
62. Kuboki, K. *et al.* Regulation of endothelial constitutive nitric oxide synthase gene expression in endothelial cells and in vivo a specific vascular action of insulin. *Circulation* **101**, 676–681 (2000).
63. Glogowski, E. A., Tsiani, E., Zhou, X., Fantus, I. G. & Whiteside, C. High glucose alters the response of mesangial cell protein kinase C isoforms to endothelin-1. *Kidney Int.* **55**, 486–499 (1999).
64. Hempel, A. *et al.* High glucose concentrations increase endothelial cell permeability via activation of protein kinase C alpha. *Circ. Res.* **81**, 363–371 (1997).
65. Williams, B., Gallacher, B., Patel, H. & Orme, C. Glucose-induced protein kinase C activation regulates vascular permeability factor mRNA expression and peptide production by human vascular smooth muscle cells in vitro. *Diabetes* **46**, 1497–1503 (1997).
66. Studer, R. K., Craven, P. A. & DeRubertis, F. R. Role for protein kinase C in the mediation of increased fibronectin accumulation by mesangial cells grown in high-glucose medium. *Diabetes* **42**, 118–126 (1993).
67. Koya, D. *et al.* Characterization of protein kinase C beta isoform activation on the gene expression of transforming growth factor-beta, extracellular matrix components, and prostanoids in the glomeruli of diabetic rats. *J. Clin. Invest.* **100**, 115–126 (1997).
68. Craven, P. A., Studer, R. K., Felder, J., Phillips, S. & DeRubertis, F. R. Nitric oxide inhibition of transforming growth factor-beta and collagen synthesis in mesangial cells. *Diabetes* **46**, 671–681 (1997).
69. Phillips, S. L., DeRubertis, F. R. & Craven, P. A. Regulation of the laminin C1 promoter in cultured mesangial cells. *Diabetes* **48**, 2083–2089 (1999).
70. Feener, E. P. *et al.* Role of protein kinase C in glucose- and angiotensin II-induced plasminogen activator inhibitor expression. *Contrib. Nephrol.* **118**, 180–187 (1996).
71. Pieper, G. M. & Riaz-ul-Haq, J. Activation of nuclear factor-kappaB in cultured endothelial cells by increased glucose concentration: prevention by calphostin C. *Cardiovasc. Pharmacol.* **30**, 528–532 (1997).
72. Yerneni, K. K., Bai, W., Khan, B. V., Medford, R. M. & Natarajan, R. Hyperglycemia-induced activation of nuclear transcription factor kappaB in vascular smooth muscle cells. *Diabetes* **48**, 855–864 (1999).
73. Koya, D. *et al.* Amelioration of accelerated diabetic mesangial expansion by treatment with a PKC beta inhibitor in diabetic db/db mice, a rodent model for type 2 diabetes. *FASEB J.* **14**, 439–447 (2000).
74. Kolm-Litty, V., Sauer, U., Nerlich, A., Lehmann, R. & Schleicher, E. D. High glucose-induced transforming growth factor beta1 production is mediated by the hexosamine pathway in porcine glomerular mesangial cells. *J. Clin. Invest.* **101**, 160–169 (1998).
75. Marshall, S., Bacote, V. & Traxinger, R. R. Discovery of a metabolic pathway mediating glucose-induced desensitization of the glucose transport system. Role of hexosamine biosynthesis in the induction of insulin resistance. *J. Biol. Chem.* **266**, 4706–4712 (1991).
76. Hawkins, M. *et al.* Role of the glucosamine pathway in fat-induced insulin resistance. *J. Clin. Invest.* **99**, 2173–2182 (1997).
77. Chen, Y. Q. *et al.* Sp1 sites mediate activation of the plasminogen activator inhibitor-1 promoter by glucose in vascular smooth muscle cells. *J. Biol. Chem.* **273**, 8225–8231 (1998).
78. Goldberg, H. J., Scholey, J. & Fantus, I. G. Glucosamine activates the plasminogen activator inhibitor 1 gene promoter through Sp1 DNA binding sites in glomerular mesangial cells. *Diabetes* **49**, 863–871 (2000).
79. Kadonaga, J. T., Courey, A. J., Ladika, J. & Tjian, R. Distinct regions of Sp1 modulate DNA binding and transcriptional activation. *Science* **242**, 1566–1570 (1988).
80. Haltiwanger, R. S., Grove, K. & Phillipsberg, G. A. Modulation of O-linked N-acetylglucosamine levels on nuclear and cytoplasmic proteins in vivo using the peptide O-GlcNAc-beta-N-acetylglucosaminidase inhibitor O-(2-acetamido-2-deoxy-D-glucopyranosylidene)amino-N-phenylcarbamate. *J. Biol. Chem.* **273**, 3611–3617 (1998).
81. Hart, G. W. Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins. *Annu. Rev. Biochem.* **66**, 315–335 (1997).
82. Du, X. D. *et al.* Hyperglycemia inhibits endothelial nitric oxide synthase activity by posttranslational modification at the AKT site. *J. Clin. Invest.* (in the press).
83. Lee, A. Y., Chung, S. K. & Chung, S. S. Demonstration that polyol accumulation is responsible for diabetic cataract by the use of transgenic mice expressing the aldose reductase gene in the lens. *Proc. Natl Acad. Sci. USA* **92**, 2780–2784 (1995).
84. Nishikawa, T. *et al.* Normalizing mitochondrial superoxide production blocks three pathways of hyperglycaemic damage. *Nature* **404**, 787–790 (2000).
85. Giugliano, D., Ceriello, A. & Paolisso, G. Oxidative stress and diabetic vascular complications. *Diabetes Care* **19**, 257–267 (1996).
86. Giardino, I., Edelstein, D. & Brownlee, M. BCL-2 expression or antioxidants prevent hyperglycemia-induced formation of intracellular advanced glycation endproducts in bovine endothelial cells. *J. Clin. Invest.* **97**, 1422–1428 (1996).
87. Korshunov, S. S., Skulachev, V. P. & Starkov, A. A. High protonic potential actuates a mechanism of production of reactive oxygen species in mitochondria. *FEBS Lett.* **416**, 15–18 (1997).
88. Craven, R. P., Phillip, S. L., Melhem, M. F., Liachenko, J. & DeRubertis, F. R. Overexpression of Mn²⁺ superoxide dismutase suppresses increases in collagen accumulation induced by culture in mesangial cells in high media glucose. *Metabolism* (in the press).
89. Yamagishi, S. I., Edelstein, D., Du, X. L. & Brownlee, M. Hyperglycemia potentiates collagen-induced platelet activation through mitochondrial superoxide overproduction. *Diabetes* **50**, 1491–1494 (2001).
90. Craven, P. A., Melham, M. F., Phillip, S. L. & DeRubertis, F. R. Overexpression of Cu²⁺/Zn²⁺ superoxide dismutase protects against early diabetic glomerular injury in transgenic mice. *Diabetes* **50**, 2114–2125 (2001).
91. Engerman, R. L. & Kern, T. S. Progression of incipient diabetic retinopathy during good glycemic control. *Diabetes* **36**, 808–812 (1987).
92. The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group. Retinopathy and nephropathy in patients with type 1 diabetes four years after a trial of intensive therapy. *N. Engl. J. Med.* **342**, 381–389 (2000).
93. Quinn, M., Angelico, M. C., Warram, J. H. & Krolewski, A. S. Familial factors determine the development of diabetic nephropathy in patients with IDDM. *Diabetologia* **39**, 940–945 (1996).
94. The Diabetes Control and Complications Trial Research Group. Clustering of long-term complications in families with diabetes in the diabetes control and complications trial. *Diabetes* **46**, 1829–1839 (1997).
95. Wagenknecht, L. E. *et al.* Familial aggregation of coronary artery calcium in families with type 2 diabetes. *Diabetes* **50**, 861–866 (2001).
96. Kowluru, R. A., Tang, J. & Kern, T. S. Abnormalities of retinal metabolism in diabetes and experimental galactosemia. VII. Effect of long-term administration of antioxidants on the development of retinopathy. *Diabetes* **50**, 1938–1942 (2001).
97. Ting, H. H. *et al.* Vitamin C improves endothelium-dependent vasodilation in patients with non-insulin-dependent diabetes mellitus. *J. Clin. Invest.* **97**, 22–28 (1996).
98. Heart Outcomes Prevention Evaluation Study Investigators. Effects of ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus: results of the HOPE study and MICRO-HOPE substudy. *Lancet* **355**, 253–259 (2000).
99. Salvemini, D. *et al.* A nonpeptidyl mimic of superoxide dismutase with therapeutic activity in rats. *Science* **286**, 304–306 (1999).
100. Coppey, L. J. *et al.* *Brit. J. Pharmacol.* **134**, 21–29 (2001).

Acknowledgements

This work was supported by grants from NIH, Juvenile Diabetes Research Foundation and American Diabetes Association. Owing to space limitations, a comprehensive list of reference citations could not be included. I apologize to those colleagues whose work is not specifically referenced, and gratefully acknowledge their contributions to the field.

New drug targets for type 2 diabetes and the metabolic syndrome

David E. Moller

Departments of Molecular Endocrinology and Metabolic Disorders, Merck Research Laboratories, Rahway, New Jersey 07065, USA

An insidious increase in features of the 'metabolic syndrome' — obesity, insulin resistance and dyslipidaemia — has conspired to produce a worldwide epidemic of type 2 insulin-resistant diabetes mellitus. Most current therapies for this disease were developed in the absence of defined molecular targets or an understanding of disease pathogenesis. Emerging knowledge of key pathogenic mechanisms, such as the impairment of glucose-stimulated insulin secretion and the role of 'lipotoxicity' as a probable cause of hepatic and muscle resistance to insulin's effects on glucose metabolism, has led to a host of new molecular drug targets. Several have been validated through genetic engineering in mice or the preliminary use of lead compounds and therapeutic agents in animals and humans.

Type 2 insulin-resistant diabetes mellitus accounts for 90–95% of all diabetes. This heterogeneous disorder afflicts an estimated 6% of the adult population in Western society; its worldwide frequency is expected to continue to grow by 6% per annum, potentially reaching a total of 200–300 million cases in 2010 (refs 1, 2). The main force driving this increasing incidence is a staggering increase in obesity, the single most important contributor to the pathogenesis of diabetes.

It is now clear that aggressive control of hyperglycaemia in patients with type 2 diabetes can attenuate the development of chronic complications such as retinopathy and nephropathy³. At present, therapy for type 2 diabetes relies mainly on several approaches intended to reduce the hyperglycaemia itself: sulphonylureas (and related insulin secretagogues), which increase insulin release from pancreatic islets; metformin, which acts to reduce hepatic glucose production; peroxisome proliferator-activated receptor- γ (PPAR γ) agonists (thiazolidinediones), which enhance insulin action; α -glucosidase inhibitors, which interfere with gut glucose absorption; and insulin itself, which suppresses glucose production and augments glucose utilization (Table 1). These therapies have limited efficacy, limited tolerability and significant mechanism-based side effects. Of particular concern is the tendency for most treatments to enhance weight gain. Several current approaches are also associated with episodes of hypoglycaemia, and few of the available therapies adequately address underlying defects such as obesity and/or insulin resistance. A problem particular to the sulphonylureas is that many patients who respond initially become refractory to treatment over time ('secondary failures'). Thus, newer approaches are desperately needed. Particular emphasis should be placed on finding and using mechanisms that are dependent on physiological responses (for example, glucose-mediated insulin secretagogues), and that result in weight loss (or lack of weight gain).

'Metabolic syndrome' encompasses type 2 diabetes (or prediabetes) and a common constellation of closely linked clinical features⁴. Characteristic factors include insulin resistance *per se*, obesity (in particular abdominal adiposity), hypertension, and a common form of dyslipidaemia (raised triglycerides and low high-density lipoprotein (HDL)-cholesterol with or without elevation of low-density lipoprotein (LDL)-cholesterol). Metabolic syndrome is

associated with a markedly increased incidence of coronary, cerebral and peripheral artery disease. Thus, atherosclerotic cardiovascular disease (ASCVD) is responsible for 80% of diabetic mortality and more than 75% of all hospitalizations for diabetic complications. Indeed, type 2 diabetes now represents a coronary heart disease 'risk equivalent'; this means that the risk of myocardial infarction in patients with diabetes and no history of cardiac disease roughly equates to the risk in non-diabetic patients with known cardiac disease⁵. Put another way, for any abnormality in risk factors, diabetics have a two- to fourfold greater ASCVD risk than do people without diabetes. Thus, therapeutic approaches that not only lower glucose, but also specifically address diabetic dyslipidaemia and ASCVD complications are critically needed. In addition, current therapies do not directly address the other late-stage complications of diabetes (for example, neuropathy and retinopathy) that constitute a major disease burden (see the review in this issue by Brownlee, pages 813–820, for further discussion).

As therapeutic approaches for type 2 diabetes continue to evolve and improve, the goal of future treatment will be to intervene when very early clinical signs, such as impaired glucose tolerance and other aspects of metabolic syndrome, first manifest. The availability of drugs that affect underlying mechanisms may lead to a new therapeutic paradigm for the prevention of diabetes and its complications.

Current therapeutic approaches were largely developed in the absence of defined molecular targets or even a solid understanding of disease pathogenesis. Within the past few years, our understanding of biochemical pathways related to the

Table 1 Current therapeutic agents for type 2 diabetes

Drug class	Molecular target	Site(s) of action	Adverse events
Insulin	Insulin receptor	Liver, muscle, fat	Hypoglycaemia, weight gain
Sulphonylureas (e.g. glibenclamide) plus nateglinide and repaglinide	SU receptor/ K ⁺ ATP channel	Pancreatic β -cell	Hypoglycaemia, weight gain
Metformin — biguanides	Unknown	Liver (muscle)	Gastrointestinal disturbances, lactic acidosis
Acarbose	α -glucosidase	Intestine	Gastrointestinal disturbances
Pioglitazone, rosiglitazone (thiazolidinediones)	PPAR γ	Fat, muscle, liver	Weight gain, oedema, anaemia

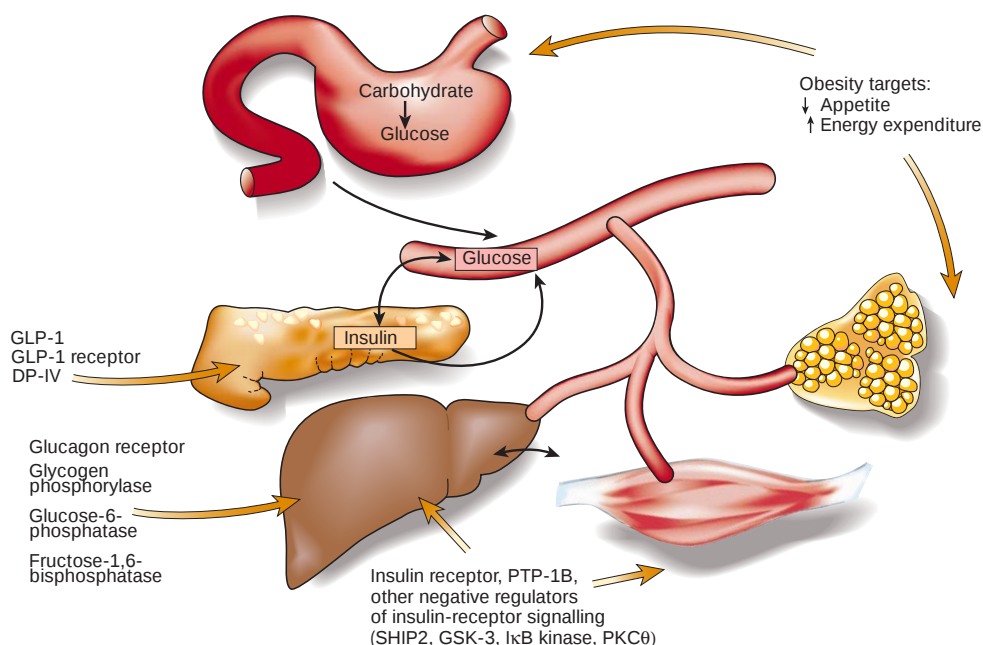


Figure 1 A better understanding of defects involving several key organ systems has led to new drug targets for type 2 diabetes. The liver is largely responsible for unrestrained glucose production through increased rates of gluconeogenesis and glycogenolysis. Potential drug targets that modulate these processes include the glucagon receptor (antagonists), glycogen phosphorylase (inhibitors), and other rate-controlling enzymes such as glucose-6-phosphatase and fructose-1,6-bisphosphatase (inhibitors). Defective glucose-stimulated insulin secretion by pancreatic islet β -cells could be alleviated with recombinant glucagon-like peptide 1 (GLP-1) or agonists of the GLP-1

receptor. Alternatively, decreased GLP-1 clearance can be achieved with inhibition of dipeptidylpeptidase IV (DP-IV). To reduce insulin resistance, enhanced insulin action in liver and muscle (and fat) might be achieved with small-molecule activators of the insulin receptor or inhibitors of protein tyrosine phosphatase (PTP)-1B. Other potential drug targets in the insulin signalling pathway are discussed in the text. The development of anti-obesity agents that produce reduced appetite and/or increased energy expenditure will also lead to effective treatment (and prevention) of type 2 diabetes.

development of metabolic syndrome has expanded. There is an unprecedented range of molecular drug targets within these pathways. They have been identified on the basis of predicted roles in modulating one or more key aspects of the pathogenesis of diabetes and metabolic syndrome. Several mechanistic categories for new therapeutic approaches can be considered. First are approaches aimed at reducing excessive glucose production by the liver; second, mechanisms to augment glucose-stimulated insulin secretion; third, specific molecular targets in the insulin signalling pathway; and fourth, new approaches to obesity and altered lipid metabolism, which offer the prospect of net improvements in insulin action (or secretion) (Fig. 1).

Reducing excessive hepatic glucose production

The liver has a critical role in regulating endogenous glucose production from *de novo* synthesis (gluconeogenesis) or the catabolism of glycogen (glycogenolysis). Increased rates of hepatic glucose production are largely responsible for the development of overt hyperglycaemia, in particular fasting hyperglycaemia, in patients with diabetes⁶. A relative decrease in insulin levels, or reduced hepatic responsiveness to insulin, can lead to increased output of glucose by the liver. Several drug targets in the liver offer new ways of attenuating excessive hepatic glucose production.

Glucagon is a well described hormone that contributes to hyperglycaemia through the induction of both gluconeogenic and glycogenolytic pathways⁷⁻⁹. The glucagon receptor, a seven-transmembrane domain G-protein-coupled receptor, is an obvious target for the development of small-molecule antagonists¹⁰. Neutralizing antibodies¹¹ and peptide-receptor antagonists¹² have both been shown to be effective antagonists *in vivo*. Several non-peptide glucagon-receptor antagonists have been reported so far¹⁰; although only one has shown hints of efficacy in early-stage human clinical trials¹³.

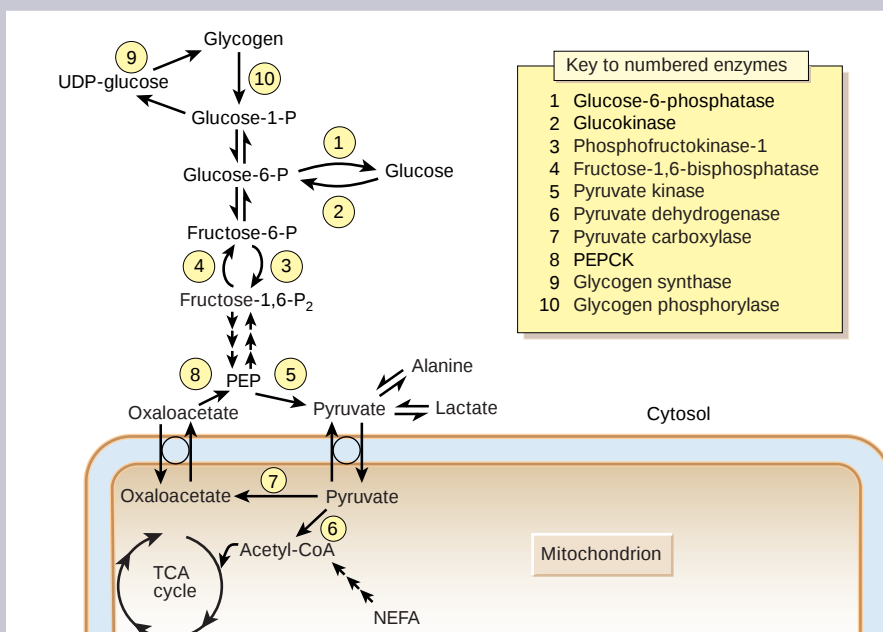
Beyond the control of glucagon action, several enzymes that regulate rate-controlling steps in the gluconeogenic or glycogenolytic pathways are obvious molecular targets for therapeutic intervention (Fig. 2). An approach that seems to have advanced into early-stage clinical trials entails inhibition of hepatic glycogen phosphorylase¹⁴, an enzyme that catalyses the release of monomeric glucose from stored glycogen. Despite some evidence that glycogenolysis may account for only a small fraction of liver glucose production in type 2 diabetes¹⁵, an inhibitory molecule (a chloroindole carboxamide compound) that binds to a novel allosteric site (distinct from the enzyme active site) on glycogen phosphorylase has been highly effective in rodent models of diabetes¹⁴. The potential of such inhibitors to impair exercise-mediated catabolism of muscle glycogen is, however, a concern that merits further study.

Other hepatic enzyme targets that have received more limited attention include fructose-1,6-bisphosphatase and glucose-6-phosphatase^{16,17}. Inhibition of the former would selectively block gluconeogenesis by disrupting the conversion of fructose-1,6-bisphosphate to fructose-6-phosphate. Inhibition of glucose-6-phosphatase would attenuate the final step in hepatic glucose production common to both gluconeogenic and glycogenolytic pathways. Although inhibition of hepatic glucose production remains attractive for further study, there are several potential liabilities inherent in this approach, including hypoglycaemia, accumulation of hepatic triglycerides and increased plasma lactate levels.

Enhancing glucose-stimulated insulin secretion

A key component of the pathophysiology of type 2 diabetes involves a relatively selective defect in the ability of glucose to provoke secretion of insulin from pancreatic islet β -cells (see review in this issue by Mathis, Vence and Benoist, pages 792–798). This defect accounts for

Figure 2 Important pathways regulating glucose metabolism in the liver. Excessive hepatic glucose output occurs in diabetes through increases in glycogenolysis and/or gluconeogenesis. Inhibitors of glycogen phosphorylase inhibit glucose output by decreasing hepatic glycogen catabolism. Other relevant targets include fructose-1,6-bisphosphatase, which controls a rate-limiting step in gluconeogenesis, and glucose-6-phosphatase, which catalyses the final common step required for release of glucose from the liver. NEFA, non-essential fatty acids; PEP, phosphoenolpyruvate. [Adapted from ref. 77 with permission.]



the failure of the β -cell to compensate for increasing insulin resistance and for the ultimate development of overt hyperglycaemia. In addition, ample evidence shows that defective β -cell function can be an early predisposing factor¹⁸. Unlike sulphonylureas and related compounds, which stimulate insulin secretion in the absence of high glucose levels and work by blocking ATP-sensitive K^+ channels, more desirable alternative approaches would potentiate insulin secretion in a purely glucose-dependent fashion.

In this regard, two distinct gut-derived peptide hormones — glucagon-like peptide-1 (GLP-1) and gastric inhibitory peptide (GIP) — act through their respective G-protein-coupled receptors on β -cells to potentiate glucose-stimulated insulin secretion¹⁹. Administration of either hormone to humans can potentiate insulin secretion, whereas selective gene disruptions of either the GLP-1 or GIP receptors produces a phenotype of impaired glucose-stimulated insulin secretion^{19,20}. Additional mechanisms by which GLP-1 could have anti-diabetic effects include inhibition of gastric emptying, impairment of glucagon secretion and potential central anorectic effects¹⁹. Moreover, cell-based and animal studies suggest that GLP-1 has the potential to promote the growth of new islets and β -cell hyperplasia¹⁹. Administration of either exogenous GLP-1 or a potent GLP-1 agonist (exendin 4) derived from lizard venom has been shown to suppress energy intake in humans (reviewed in ref. 19), providing even stronger validation of GLP-1 as a therapeutic agent. A potential anti-obesity effect can thus be envisaged. In contrast, infusion of a GLP-1 antagonist (exendin 9-39 amide) impaired post-prandial glucose control²¹.

Although both GLP-1 (and GIP) have strong potential as chronic therapies for diabetes, both are subject to rapid amino-terminal degradation ($t_{1/2} \sim 1$ min) by dipeptidylpeptidase-IV (DP-IV, also known as CD26), a proline-specific serine dipeptidase. GLP-1 thus becomes inactivated by DP-IV *in vitro* ($k_{cat}/K_m \sim 1 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$) to generate GLP-1[9-36] (Fig. 3). One approach to this problem could be the use of modified GLP-1 peptide agonists that are resistant to DP-IV (such as exendin 4). Importantly, specific DP-IV inhibitors have also been shown to increase circulating GLP-1 in both rodents and humans. Validation of DP-IV as a relevant drug target was bolstered by observations that DP-IV-null mice have increased circulating active GLP-1[7-36] along with enhanced insulin secretion and an otherwise healthy phenotype²². Moreover, early-stage clinical trials have provided 'proof-of-concept' for efficacy of DP-IV inhibition in humans with type 2 diabetes^{23,24}.

Although small-molecule agonists of the GLP-1 (and GIP) receptors would seem to provide a logical drug option, the existence of viable lead compounds has yet to be reported. The further development of GLP-1 analogues and DP-IV inhibitors is, however, likely to yield important new therapeutic approaches that might circumvent the liabilities of hypoglycaemia, weight gain and secondary failures associated with sulphonylurea use.

Targeting the insulin signalling pathway

The role of peripheral and hepatic insulin resistance in the pathogenesis of diabetes is undisputed. As discussed by Saltiel and Kahn in an accompanying review (pages 799–806), insulin resistance can be due to multiple defects in signal transduction (such as impaired activation of insulin receptor-tyrosine kinase and reduced activation of insulin-stimulated phosphatidylinositol-3-OH kinase (PI(3)K)). A number of molecular targets are now being investigated as ways of enhancing insulin-mediated signal transduction (Table 2).

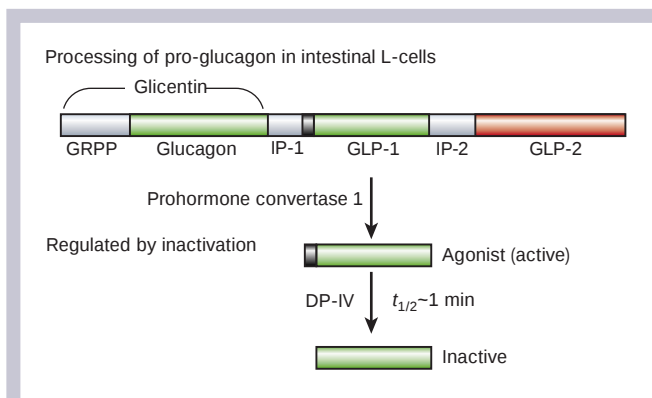


Figure 3 Biosynthesis and regulation of glucagon-like peptide 1 (GLP-1). GLP-1 is a product of the pre-pro-glucagon gene. Pro-glucagon is cleaved by prohormone convertase 1 to generate active GLP-1[7-36], which is released from intestinal L-cells during nutrient ingestion. GLP-1 is rapidly hydrolysed *in vivo* ($t_{1/2} \sim 1$ min) to produce an inactive product, GLP-1[9-36]. DP-IV, a proline-specific serine dipeptidase, is solely responsible for this inactivation. DP-IV inhibitors therefore represent an indirect therapeutic approach to stabilizing endogenous GLP-1.

Non-peptide small molecules that can activate the insulin receptor, or potentiate its activation by insulin, have proved elusive. But the recent discovery of a small-molecule natural-product derivative that mediates selective activation of the insulin receptor²⁵ is encouraging.

An alternative approach to targeting the insulin receptor itself would be to inhibit enzymes responsible for deactivation of the receptor or downstream targets in the signalling pathway (for example, IRS proteins). A number of specific protein tyrosine phosphatases (PTPs) have been identified as candidate targets²⁶. Vanadium, peroxovanadium and their derivatives are non-selective PTP inhibitors. Nevertheless, demonstration of the insulin-sensitizing efficacy of vanadyl sulphate in humans suggests that one or more PTPs may be viable drug targets²⁷. PTP-1B is an intracellular enzyme specifically implicated in the negative regulation of insulin signalling²⁶. Recent results from genetic knockout of PTP-1B provide strong validation of this particular PTP as a potential target. The PTP-1B-null mice are healthy and have markedly enhanced sensitivity to insulin²⁸. Surprisingly, they also showed substantial resistance to diet-induced obesity²⁸. This observation is hard to reconcile with the fact that insulin serves as a major anabolic hormone to potentiate adipose accretion. Insulin action in the brain, however, may enhance satiety and mediate increased energy expenditure. Further validation of PTP-1B as a drug target has been provided by evidence of increased insulin action in insulin-resistant rats treated with a PTP-1B antisense oligonucleotide²⁹. This treatment seems to have worked with injections of the antisense oligonucleotide once to twice weekly, and could be a viable approach for testing in humans.

Other putative negative regulators of insulin signalling (Table 2) have recently been implicated as independent drug targets. Glycogen synthase kinase-3 (GSK-3) has a clear role in opposing the effect of insulin, by inhibiting the activation of glycogen synthase and the subsequent accumulation of glycogen in muscle³⁰. Recent results with potent and selective inhibitors suggest that reducing GSK-3 activity *in vivo* could indeed augment insulin action, and that this may occur at multiple steps³¹. This serine-threonine kinase may, however, have an important role in the regulation of cell proliferation and apoptosis through its function within the Wnt signalling pathway³⁰.

SH2-domain-containing inositol 5-phosphatase type 2 (SHIP2) may function to dephosphorylate key phospholipids (for example, phosphatidylinositol-3-phosphate; PtdIns(3)P) that are generated by insulin-mediated PI(3)K activation (see the review in this issue by Saltiel and Kahn, pages 799–806). This enzyme was implicated recently as a diabetes target, as heterozygous null mice have markedly enhanced sensitivity to insulin³².

Table 2 Potential drug targets in the insulin signalling pathway		
Target	Validation	Potential mechanism(s)
Insulin receptor	Insulin, small-molecule activators/potentiators	Apparent direct activation of the receptor
PTP-1B	Efficacy of vanadium compounds; PTP-1B ^{-/-} (null) mice (insulin sensitive and obesity resistant); efficacy of PTP-1B antisense oligonucleotide	Mediates dephosphorylation of the insulin receptor (and its tyrosyl-phosphorylated substrates)
SHIP-2	SHIP-2 ^{-/-} mice (insulin sensitive)	Dephosphorylation of phosphoinositides (for example, products of PI(3)K)
GSK-3	Efficacy of GSK-3 inhibitors in rodent models	Phosphorylation of glycogen synthase leading to inhibition of glycogen synthesis; potential negative regulation of other insulin signalling events
IκB kinase	Efficacy of high-dose salicylate (inhibits IκB kinase); IκB kinase ^{-/-} mice (insulin sensitive)	Serine-threonine phosphorylation of insulin signalling intermediates (for example, IRS proteins)
PKCθ	Activated in muscle in association with fatty-acid-induced insulin resistance	Negative regulation of insulin signalling; potential serine-threonine phosphorylation of IRS proteins

Shoelson and colleagues recently revealed a new and intriguing potential diabetes drug target^{33,34}. Two lines of evidence indicate an important role for IκB kinase (IKK) as a mediator of increased protein serine or threonine phosphorylation, which has the potential to downregulate insulin signalling. First, high-dose salicylate can produce increased insulin sensitivity in association with IKK inhibition, and second, heterozygous IKK-null mice have a phenotype of increased insulin sensitivity. As tumour necrosis factor-α (TNF-α), a potential mediator of obesity-associated insulin resistance³⁵, can activate the IKK complex, a specific role for IKK in TNFα-mediated insulin resistance may be implicated³³. Similarly, protein kinase C-θ (PKCθ) could be an additional drug target, as increased muscle PKCθ activity has been observed in the context of fatty-acid-induced insulin resistance³⁶. Further elucidation of key regulators in the insulin signalling pathway will no doubt lead to the discovery of additional drug targets for type 2 diabetes.

Targeting obesity, lipid metabolism and 'lipotoxicity'

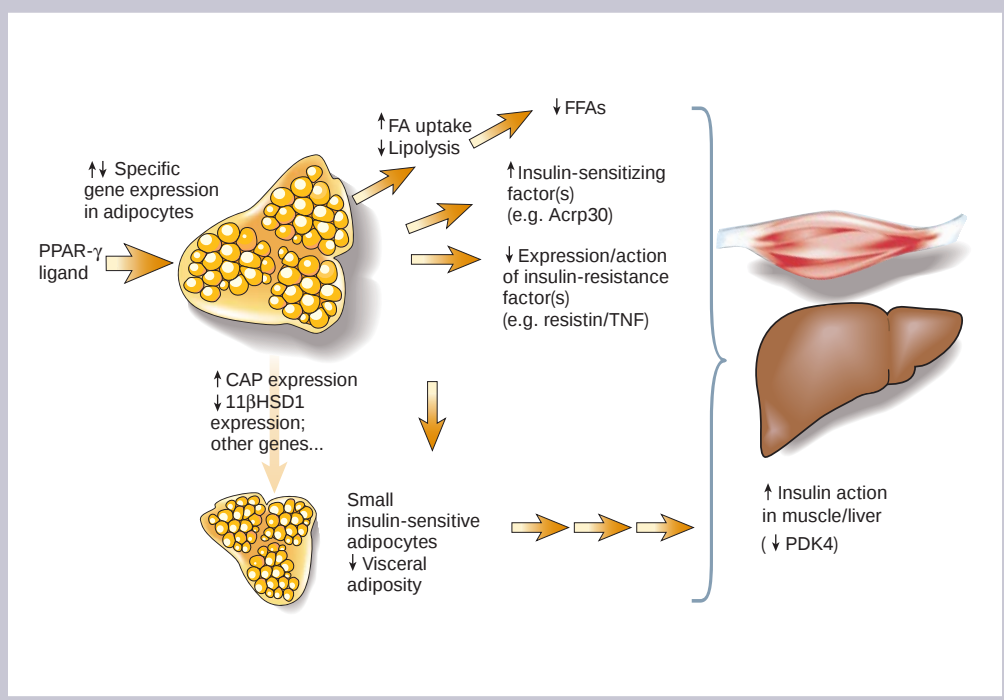
Given the critical role of obesity in the development of insulin resistance, and other features of the metabolic syndrome (see the reviews in this issue by Zimmet, Alberti and Shaw, pages 782–787, and Bell and Polonsky, 788–791), successful approaches to attenuating appetite and/or enhancing energy expenditure will prove of great benefit in preventing and treating type 2 diabetes. A wide range of drug targets for obesity *per se* is being actively investigated^{37,38}. As an example of the potential benefits of modulating one such target, agonists of the melanocortin-4 receptor (MCR-4) offer the prospect of ameliorating obesity and type 2 diabetes. Thus, either an increase in the expression of a natural MCR-4 antagonist (Agouti) or knockout of the receptor itself produces a strong phenotype with multiple features of the metabolic syndrome^{39,40}. An exciting new development⁴¹ involves the realization that approaches designed to modulate central neuroendocrine systems that control energy metabolism, such as the melanocortin pathway, can have selective and beneficial effects on peripheral metabolism (for example, to alter insulin's effect on the liver *per se*). The prospect of a new approach to appetite reduction through central inhibition of fatty-acid synthase also offers an intriguing new avenue for drug research in this area⁴².

Abnormalities of fatty-acid metabolism are increasingly recognized as key components of the pathogenesis of the metabolic syndrome and type 2 diabetes⁴³. Fat-feeding and raised levels of circulating free fatty acids (FFAs) are clearly sufficient to induce peripheral and hepatic insulin resistance⁴⁴. Accumulation of lipids inside muscle cells⁴⁵ and specific increases in muscle long-chain fatty-acyl-CoA content⁴⁴ have been implicated in causing insulin resistance. In addition, lipid accumulation within pancreatic islets has been proposed to impair insulin secretion⁴⁶. A critical player in potentiating the promoting effect of hyperinsulinaemia on hepatic lipid accumulation is the anabolic transcription factor SREBP-1, which upregulates genes such as that for fatty-acid synthase⁴⁷. These observations support a unified 'lipotoxicity' hypothesis, which states that metabolic syndrome and type 2 diabetes can be caused by the accumulation of triglycerides and long-chain fatty-acyl-CoA in liver and muscle (leading to a reduction in insulin-mediated metabolic activity) and in the islet (leading to impaired insulin secretion). Indeed, the weight-reducing hormone leptin may prevent diet-induced diabetes in rodents predominantly through reducing fat accumulation or 'steatosis' in these key tissues⁴⁸. Several therapeutic targets discussed below, including AMP-activated protein kinase (AMPK), acetyl-CoA carboxylase (ACC), adipocyte-related complement protein 30 (Acrp30), PPARγ and PPARα, represent mechanisms that could be exploited to reverse or prevent obesity-related lipotoxicity.

AMP-activated kinase and acetyl-CoA carboxylase

AMPK is activated in response to reduced cellular energy charge⁴⁹. In turn, ACC, a key AMPK substrate, is inactivated in response to phosphorylation. As ACC catalyses the formation of malonyl-CoA, a

Figure 4 Potential mechanisms of insulin sensitization by PPAR- γ ligands. The receptor PPAR- γ is predominantly expressed in adipose tissue. Ligand interactions with the receptor mediate specific changes in adipose gene expression. Altered expression of adipose genes such as fatty-acid transporter 1 may contribute to reduced production of free fatty acids (FFAs), which, in turn, is predicted to have insulin-sensitizing effects in muscle and liver. Changes in expression of other genes such as CAP or 11 β HSD1 may contribute to locally increased insulin action in adipose tissue and/or reduced visceral adiposity. Altered expression of circulating factors including TNF- α , resistin and Acrp30 is also likely to indirectly mediate increased action of insulin in liver or muscle and glucose utilization; suppression of PDK4 activity in muscle is an example of one (probably indirect) effect.



potent inhibitor of fatty-acid oxidation and the first step in fatty-acid synthesis, AMPK activation and consequent ACC inactivation will result in reduced lipid synthesis and increased fat oxidation. AMPK activation also leads to reduced hepatic SREBP-1 and to suppressed expression of downstream lipogenic genes^{49,50}. In addition, AMPK activation comprises a mechanism for exercise-mediated muscle glucose uptake^{51,52}. Allosteric AMPK activation with an adenosine analogue, AICAR, has been shown to produce several beneficial metabolic effects, including inhibition of hepatic glucose output and increased muscle glucose uptake⁴⁹. More recent results also suggest that the ability of the widely used drug metformin to inhibit hepatic glucose production and attenuate hepatic steatosis involves AMPK activation⁵⁰. Taken together, the beneficial effects of AMPK activation by AICAR, and the probable role of AMPK activation as a contributor to the metabolic effects of metformin, provide a strong rationale for targeting AMP activation as a new therapeutic approach. ACC is also an exciting independent drug target, as mice deficient in one of two main ACC isoforms (ACC2) were characterized by increased fatty-acid oxidation with markedly reduced body weight and adiposity⁵³. The extent to which ACC inhibition might protect against diabetes has not, however, been fully assessed.

Adipocyte complement-related protein 30

Acrp30, or adiponectin, a secreted adipocyte-specific protein of M_r 30,000, has recently been shown to produce beneficial metabolic effects in mice, including the ability to reduce glucose, triglycerides and FFAs⁵⁴⁻⁵⁶. This putative hormone may also induce tissue fatty-acid oxidation and reduce tissue steatosis in insulin-resistant animal models^{54,56}. An additional acute effect in enhancing hepatic insulin action has also been suggested⁵⁵. Therefore, either recombinant Acrp30 derivatives or small-molecule Acrp30-mimetic compounds could be envisaged as new therapeutic approaches.

PPARs present multiple therapeutic targets

PPARs are ligand-activated transcription factors (members of the nuclear receptor family) which offer a promising therapeutic approach to the metabolic syndrome. The known beneficial effects of PPAR ligands are largely consistent with mechanisms that can ameliorate lipotoxicity. PPAR- γ is the predominant molecular target for insulin-sensitizing thiazolidinedione (TZD) drugs^{57,58}. Surprisingly,

the TZD drug class was discovered more than a decade before⁵⁹ this mechanism was deduced⁶⁰. As TZDs suffer from only a modest net efficacy and several side effects, many investigators have been engaged in the search for improved PPAR- γ ligands as potential drugs. New compounds with markedly enhanced potency and selectivity for the receptor have recently been discovered⁵⁷.

A prevailing hypothesis for regulation of insulin sensitivity by PPAR- γ involves primary effects of PPAR- γ on gene transcription in adipose tissue (where it is most abundantly expressed), which ultimately lead to improved insulin action in muscle and liver (Fig. 4). Direct activation of PPAR- γ leads to the induction of adipocyte genes such as those for lipoprotein lipase and fatty-acid transporter 1, which in turn contribute to lowering triglyceride and FFA levels, respectively⁶¹. Similarly, suppression of TNF- α gene expression by PPAR- γ in adipose tissue has been reported⁶¹. As FFAs and TNF- α are both potential systemic mediators of insulin resistance, such effects are likely to contribute to the efficacy of PPAR- γ activation in increasing insulin sensitivity. As a consequence of reduced systemic lipid availability, muscle lipid levels can also be reduced⁶².

Recent attempts to elucidate the genes regulated by PPAR- γ and involved in insulin sensitivity have shed light on several areas of interest for drug discovery research. In some of the examples discussed below, putative downstream targets of PPAR- γ (Fig. 4) were identified using PCR-differential messenger RNA display, DNA microarrays and related techniques⁶³⁻⁶⁵. Resistin is a novel protein secreted by adipose tissue with the apparent ability to antagonize insulin action⁶⁴. Although adipose expression of resistin was initially reported to be suppressed by rosiglitazone, this finding has subsequently been questioned⁶⁶. Muscle expression of the gene for pyruvate dehydrogenase kinase 4 (PDK4) was suppressed by *in vivo* treatment of rats with PPAR- γ agonists⁶⁵. The net effect of inhibiting PDK4 should be to increase pyruvate dehydrogenase activity and to augment glucose utilization. In peripheral tissues, 11 β -hydroxysteroid dehydrogenase type 1 (11 β HSD1) catalyses the conversion of cortisone to the active glucocorticoid, cortisol. On the basis of the apparently insulin-sensitive phenotype of 11 β HSD1-null mice, inhibition of this enzyme has been suggested as a potential drug target for metabolic syndrome⁶⁷. Suppression of 11 β HSD1 by PPAR- γ agonists in adipose tissue would seem to provide further validation for this target⁶⁸. Similarly, *in vivo* activation of PPAR- γ in both

animals^{56,69} and humans⁷⁰ has also recently been shown to increase circulating Acip30 levels. Given the beneficial effects of recombinant Acip30 described above, this finding represents an intriguing potential mechanism for insulin sensitization and is a further incentive to pursue Acip30 as a target. PPAR γ -mediated induction of the expression of an adaptor protein, Cbl-associated protein (CAP), also suggests an exciting mechanistic link between PPAR γ and insulin sensitization of adipocytes *per se* (ref. 71, and see the review in this issue by Saltiel and Kahn, pages 799–806).

A closely related nuclear receptor, PPAR α , is the molecular target for the fibrate class of lipid-modulating drugs⁵⁷. Given the potential benefits of fibrate treatment for coronary disease, especially in patients with diabetes and the metabolic syndrome⁷², and recent observations suggesting an independent insulin-sensitizing effect of PPAR α agonists⁷³ (which could arise from reductions in muscle lipid content⁷⁴), the incorporation of additional PPAR α activity into compounds with PPAR γ agonist activity has been proposed⁷⁵. As adverse clinical effects of the TZD and fibrate drug classes are distinct, safe and effective insulin-sensitizing and lipid-altering compounds that engage both mechanisms of action as a single entity could possibly be developed.

By exploiting the fact that PPARs function like other nuclear receptors such as oestrogen receptors, it may also be assumed that compounds with tissue- or gene-specific effects could be identified. For the oestrogen receptor, binding of a partial agonist such as tamoxifen is associated with alternative receptor conformations, the potential for different profiles of receptor-associated co-factors, and selective biological responses versus classical agonists such as oestradiol⁷⁶. Although specific genes that mediate the adverse and beneficial effects of PPAR γ activation have yet to be fully characterized, an opportunity exists to determine *in vitro* profiles of full versus partial agonism and to test selected compounds *in vivo* to screen for those with an improved therapeutic index.

Future directions

Our collective knowledge base of pathways and discrete proteins that contribute to distinctive pathophysiological traits underlying the metabolic syndrome and type 2 diabetes is expanding rapidly. This momentum is fuelled by the quantum leap in potential 'players' provided by the annotated human genome sequence databases and molecular techniques such as DNA microarrays and gene knockouts, and the identification of potential disease genes in humans and model species. The examples of recently discovered drug targets described above strongly suggest that this increase in newly identified components of disease susceptibility will yield an even wider array of potential approaches for therapeutic intervention. In addition to small-molecule modulators of 'classical' receptor or enzyme targets, research could identify additional protein therapeutics, such as GLP-1 analogues, and even more novel approaches, such as antisense oligonucleotide-based therapies.

Intensive study of the mechanisms of action of older drugs has provided further validation of several recently identified drug targets. Further efforts in this direction are likely to be fruitful. Given the multifactorial nature of the genetic and environmental factors that contribute to the genesis of metabolic syndrome and type 2 diabetes, it is probable that further efforts to characterize disease 'sub-phenotypes' and specific genetic markers will translate into more selective therapies tailor-made for distinct subgroups of patients or those at risk of developing disease. These individuals may be identified on the basis of specific genotypes or more specific clinical markers of distinct physiological derangements. □

- Kopelman, P. G. & Hitman, G. A. Diabetes. Exploding Type II. *Lancet* **352**, SIV5 (1998).
- Amos, A. F., McCarty, D. J. & Zimmet, P. The rising global burden of diabetes and its complications: estimates and projections by 2010. *Diabet. Med.* **14**(Suppl. 5), S5–S85 (1997).
- UKPDS. UK prospective diabetes study 33: intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications with type 2 diabetes. *Lancet* **352**, 837–853 (1998).
- Executive summary of the third report of the National Cholesterol Education Program Expert Panel

- on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. *J. Am. Med. Assoc.* **285**, 2486–2496 (2001).
- Haffner, S. M., Lehto, S., Ronnemaa, T., Pyorala, K. & Laakso, M. Mortality from coronary heart disease in subjects with type 2 diabetes and in non-diabetic subjects with and without prior myocardial infarction. *N. Engl. J. Med.* **339**, 229–234 (1998).
- DeFronzo, R. A., Bonadonna, R. C. & Ferannini, E. Pathogenesis of NIDDM: a balanced overview. *Diabet. Care* **15**, 318–368 (1992).
- Unger, R. H. Glucagon physiology and pathophysiology. *N. Engl. J. Med.* **285**, 443–449 (1971).
- Roden, M. *et al.* The roles of insulin and glucagon in the regulation of hepatic glycogen synthesis and turnover in humans. *J. Clin. Invest.* **97**, 642–648 (1996).
- Shah, P., Vella, A., Basu, R., Schwenck, W. F. & Rizza, R. A. Lack of suppression of glucagon contributes to postprandial hyperglycemia in subjects with type 2 diabetes mellitus. *J. Clin. Endocrinol. Metab.* **85**, 4053–4059 (2000).
- Connell, R. D. Glucagon antagonists for the treatment of type 2 diabetes. *Exp. Opin. Ther. Patents* **9**, 701–709 (1999).
- Brand, C. L. *et al.* Immunoneutralization of endogenous glucagon with monoclonal glucagon antibody normalizes hyperglycemia in moderately streptozotocin-diabetic rats. *Diabetologia* **37**, 985–993 (1994).
- Unson, C. G., Andreu, D., Gurzenda, E. M. & Merrifield, R. B. Synthetic peptide antagonists of glucagon. *Proc. Natl Acad. Sci. USA* **84**, 4083–4087 (1987).
- Petersen, K., Sullivan, J., Amatruda, S. J., Livingston, J. & Shulman, G. *Diabetologia* **42**(Suppl.), A42 (1999).
- Treadway, J. L., Mendys, P. & Hoover, D. J. Glycogen phosphorylase inhibitors for the treatment of type 2 diabetes mellitus. *Exp. Opin. Invest. Drugs* **10**, 439–454 (2001).
- Magnusson, I., Rothman, D. L., Katz, I. D., Shulman, R. D. & Shulman, G. I. Increased rate of gluconeogenesis in Type II diabetes mellitus: a 13C nuclear magnetic resonance study. *J. Clin. Invest.* **90**, 1323–1327 (1992).
- Zhang, B. & Moller, D. E. New approaches in the treatment of type 2 diabetes. *Curr. Opin. Chem. Biol.* **4**, 461–467 (2000).
- Parker, J. C. *et al.* Plasma glucose levels are reduced in rats and mice treated with an inhibitor of glucose-6-phosphate translocase. *Diabetes* **47**, 1630–1636 (1998).
- Porte, D. J. Banting Lecture 1990: beta-cells in type II diabetes mellitus. *Diabetes* **40**, 166–180 (1991).
- Drucker, D. J. The glucagon-like peptides. *Endocrinology* **142**, 521–527 (2001).
- Miyawaki, K. *et al.* Glucose intolerance caused by a defect in the entero-insular axis: a study in gastric inhibitory polypeptide knockout mice. *Proc. Natl Acad. Sci. USA* **96**, 14843–14847 (1999).
- Edwards, C. M. *et al.* Glucagon-like peptide 1 has a physiological role in the control of postprandial glucose in humans. *Diabetes* **48**, 86–93 (1999).
- Marguet, D. *et al.* Enhanced insulin secretion and improved glucose tolerance in mice lacking CD26. *Proc. Natl Acad. Sci. USA* **97**, 6874–6879 (2000).
- Demuth, H.-U. *et al.* Single dose treatment of diabetic patients by the DP IV inhibitor P32/98. *Diabetes* **49**(Suppl. 1), A102 (2000).
- Ahren, B. *et al.* Inhibition of DPPIV by NVP DPP728 improves metabolic control over a 4 week period in type 2 diabetes. *Diabetes* **50**(Suppl. 2), A104 (2001).
- Zhang, B. *et al.* Discovery of a small molecule insulin mimetic with antidiabetic activity in mice. *Science* **284**, 974–977 (1999).
- Goldstein, B. J., Li, P. M., Ding, W. D., Ahmad, F. & Zhang, W. R. in *Vitamins and Hormones—Advances in Research and Applications* Vol. 54 (ed. Litwack, J.) 67–96 (Academic, San Diego, 1998).
- Cohen, N. *et al.* Oral vanadyl sulphate improves hepatic and peripheral insulin sensitivity in patients with non-insulin dependent diabetes mellitus. *J. Clin. Invest.* **95**, 2501–2509 (1995).
- Elchebly, M. *et al.* Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. *Science* **283**, 1544–1548 (1999).
- Bush, E. N. *et al.* Treatment of Zucker diabetic fatty rats with antisense oligonucleotide to phosphotyrosine phosphatase-1B for 5 weeks halts development of diabetes. *Diabetes* **50**(Suppl. 2), A81 (2001).
- Weston, C. R. & Davis, R. J. Signaling specificity—a complex affair. *Science* **292**, 2439–2440 (2001).
- Henriksen, E. J. *et al.* Glycogen synthase kinase-3 inhibitors potentiate glucose tolerance and muscle glycogen synthase activity in the Zucker Diabetic Fatty Rat. *Diabetes* **50**(Suppl. 2), A279 (2001).
- Clement, S. *et al.* The lipid phosphatase SHIP2 controls insulin sensitivity. *Nature* **409**, 92–97 (2001).
- Yuan, M. *et al.* Reversal of obesity- and diet-induced insulin resistance with salicylates or targeted disruption of Ikk β . *Science* **293**, 1673–1677 (2001).
- Kim, J. K. *et al.* Prevention of fat-induced insulin resistance by salicylate. *J. Clin. Invest.* **108**, 437–446 (2001).
- Moller, D. E. Potential role of TNF α in the pathogenesis of insulin resistance and type II diabetes. *Trends Endocrinol. Metab.* **11**, 212–217 (2000).
- Shulman, G. I. Cellular mechanisms of insulin resistance. *J. Clin. Invest.* **106**, 171–176 (2000).
- Moller, D. E. & Van der Ploeg, L. H. T. in *Handbook of Experimental Pharmacology: Obesity Pathology and Therapy* (eds Lockwood, D. & Heffner, T.) 404–426 (Springer, Berlin, 2000).
- Spiegelman, B. M. & Flier, J. S. Obesity and the regulation of energy balance. *Cell* **104**, 531–543 (2001).
- Klebig, M. L., Wilkinson, J. E., Geisler, J. G. & Woychik, R. P. Ectopic expression of the agouti gene in transgenic mice causes obesity, features of type II diabetes, and yellow fur. *Proc. Natl Acad. Sci. USA* **92**, 4728–4732 (1995).
- Huszar, D. *et al.* Targeted disruption of the melanocortin-4 receptor results in obesity. *Cell* **88**, 131–141 (1997).
- Obici, S. *et al.* Central melanocortin receptors regulate insulin action. *J. Clin. Invest.* **108**, 1079–1085 (2001).
- Loftus, T. M. *et al.* Reduced food intake and body weight in mice treated with fatty acid synthase inhibitors. *Science* **288**, 2379–2381 (2000).
- McGarry, J. D. What if Minkowski had been agueic? An alternative angle on diabetes. *Science* **258**, 766–770 (1992).
- Ruderman, N. B., Saha, A. K., Vavvas, D. & Witters, L. A. Malonyl-CoA, fuel sensing, and insulin resistance. *Am. J. Physiol.* **276**, E1–E18 (1999).
- Krassak, M. *et al.* Intramyocellular lipid concentrations are correlated with insulin sensitivity in humans: a 1H NMR spectroscopy study. *Diabetologia* **42**, 113–116 (1999).
- Unger, R. H. Lipotoxicity in the pathogenesis of obesity-dependent NIDDM. *Diabetes* **44**, 863–870 (1995).
- Shimomura, I. *et al.* Decreased IRS-2 and increased SREBP-1c lead to mixed insulin resistance and sensitivity in livers of lipodystrophic and ob/ob mice. *Mol. Cell.* **6**, 77–86 (2000).

48. Lee, Y. *et al.* Liporegulation in diet-induced obesity: the antisteatotic role of hyperleptinemia. *J. Biol. Chem.* **276**, 5629–5635 (2000).
49. Winder, W. W. & Hardie, D. G. The AMP-activated protein kinase, a metabolic master switch: possible roles in type 2 diabetes. *Am. J. Physiol.* **277**, E1–E10 (1999).
50. Zhou, G. *et al.* Role of AMP-activated protein kinase in mechanism of metformin action. *J. Clin. Invest.* **108**, 1167–1179 (2001).
51. Goodyear, L. J. AMP activated protein kinase: a critical signaling intermediary for exercise-stimulated glucose uptake? *Exercise Sport Sci. Rev.* **28**, 113–116 (2000).
52. Mu, J., Brozinick, J. T., Valladares, O., Bucan, M. & Birnbaum, M. J. A role for AMP-activated protein kinase in contraction- and hypoxia-regulated glucose transport in skeletal muscle. *Mol. Cell.* **7**, 1085–1094 (2001).
53. Abu-Elheiga, L., Matzuk, M. M., Abo-Hashema, K. A. H. & Wakil, S. J. Continuous fatty acid oxidation and reduced fat storage in mice lacking acetyl-CoA carboxylase 2. *Science* **291**, 2613–2616 (2001).
54. Fruebis, J. *et al.* Proteolytic cleavage product of 30-kDa adipocyte complement-related protein increases fatty acid oxidation in muscle and causes weight loss in mice. *Proc. Natl Acad. Sci. USA* **98**, 2005–2010 (2001).
55. Berg, A. H., Combs, T., Du, X., Brownlee, M. & Scherer, P. E. The adipocyte-secreted protein Acrp30 enhances hepatic insulin action. *Nature Med.* **7**, 947–953 (2001).
56. Yamauchi, T. *et al.* The fat-derived hormone adiponectin reverses insulin resistance associated with both lipoatrophy and obesity. *Nature Med.* **7**, 941–946 (2001).
57. Willson, T. M., Brown, P. J., Sternbach, D. D. & Henke, B. R. The PPARs: from orphan receptors to drug discovery. *J. Med. Chem.* **43**, 527–550 (2000).
58. Moller, D. E. & Greene, D. A. Peroxisome proliferator-activated receptor (PPAR) γ agonists for diabetes. *Adv. Prot. Chem. (Drug Discovery)* **56**, 181–212 (2001).
59. Sohda, T. *et al.* Studies on antidiabetic agents. Synthesis of 5-[4-(1-methylcyclohexylmethoxy) benzyl]thiazolidine-2,4-dione (AD-3878) and its derivatives. *Chem. Pharm. Bull.* **30**, 3580–3600 (1982).
60. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor γ . *J. Biol. Chem.* **270**, 12953–12956 (1995).
61. Berger, J. & Moller, D. E. Mechanism of PPAR action. *Annu. Rev. Med.* (in the press).
62. Oakes, N. D. *et al.* A new antidiabetic agent, BRL 49653, reduces lipid availability and improves insulin action and glucoregulation in the rat. *Diabetes* **43**, 1203–1210 (1994).
63. Gerhold, D. *et al.* Gene expression profile of adipocyte differentiation and its regulation by PPAR γ agonist. *Diabetes* **49**(Suppl. 1), A212 (2000).
64. Steppan, C. M. *et al.* The hormone resistin links obesity to diabetes. *Nature* **409**, 307–312 (2001).
65. Way, J. M. *et al.* Comprehensive messenger ribonucleic acid profiling reveals that peroxisome proliferator-activated receptor γ activation has coordinate effects on gene expression in multiple insulin-sensitive tissues. *Endocrinology* **142**, 1269–1277 (2001).
66. Way, J. M. *et al.* Adipose tissue resistin expression is severely suppressed in obesity and stimulated by peroxisome proliferator-activated receptor γ agonists. *J. Biol. Chem.* **276**, 25651–25653 (2001).
67. Seckl, J. R. & Walker, B. R. 11 β -hydroxysteroid dehydrogenase Type 1—a tissue-specific amplifier of glucocorticoid action. *Endocrinology* **142**, 1371–1376 (2001).
68. Berger, J. *et al.* Peroxisome proliferator-activated receptor- γ ligands inhibit adipocyte 11 β -hydroxysteroid dehydrogenase type 1 expression and activity. *J. Biol. Chem.* **276**, 12629–12635 (2001).
69. Combatsiaris, T. *et al.* Induction of Acrp30 levels by PPAR γ agonists: a potential mechanism of insulin sensitization. *Diabetes* **50**(Suppl. 2), A271 (2001).
70. Wagner, J. A. *et al.* Putative biomarkers to discriminate the PPAR α / γ effects of fenofibrate and rosiglitazone in normal volunteers. *Diabetes* **50**(Suppl. 2), A134 (2001).
71. Baumann, C. A., Chokshi, N., Saltiel, A. R. & Ribon, V. Cloning and characterization of a functional peroxisome proliferator activator receptor- γ -responsive element in the promoter of the CAP gene. *J. Biol. Chem.* **275**, 9131–9135 (2000).
72. Rubins, H. B. & Robins, S. J. Conclusions from the VA-HIT study. *Am. J. Cardiol.* **86**, 543–544 (2000).
73. Guerre-Millo, M. *et al.* Peroxisome proliferator-activated receptor α activators improve insulin sensitivity and reduce adiposity. *J. Biol. Chem.* **275**, 16638–16642 (2000).
74. Ye, J.-M. *et al.* Peroxisome proliferator-activated receptor- α activation lowers muscle lipids and improves insulin sensitivity in high fat-fed rats. *Diabetes* **50**(Suppl. 2), 411–417 (2001).
75. Murakami, K. *et al.* A novel insulin sensitizer acts as a coligand for peroxisome proliferator-activated receptor- α (PPAR- α) and PPAR- γ : effect of PPAR- α activation on abnormal lipid metabolism in liver of Zucker fatty rats. *Diabetes* **47**, 1841–1847 (1998).
76. McDonnell, D. P., Clemm, D. L., Hermann, T., Goldman, M. E. & Pike, J. W. Analysis of estrogen receptor function *in vitro* reveals three distinct classes of antiestrogens. *Mol. Endocrinol.* **9**, 659–669 (1995).
77. Pickup, J. & Williams, G. (eds) *Textbook of Diabetes* 2nd edn Vol. 1 Ch. 11 p.11.7 (Blackwell Science, London, 1997).

Acknowledgements

I thank many colleagues at Merck Research Laboratories who have contributed to ideas discussed in this review. In particular, I also thank N. Thornberry, L. Rossetti and B. Zhang for helpful reviews of the manuscript. I apologize to numerous colleagues whose original papers were not cited owing to space limitations.

BRISTOL-MYERS SQUIBB AND DIABETES:

Basic Science, Clinical Development, Global Surveillance



The future of medicine has never been as bright as it is today. As a result of advances in technology, which have made the sequencing of the human genome possible, we stand at the brink of a golden age of medical research. Yet, despite the promises of these giant steps forward, we still wrestle with the major unmet medical needs of today. One such unmet need spurs our ongoing research into new medicines to fight diabetes.



Many people first become aware that they have diabetes by developing one of its life-threatening complications, which include blindness, kidney disease, nerve disease, heart disease and stroke. By then, it may be too late. It is vital to quickly diagnose and treat diabetes. Improved diet and increased exercise may improve the quality of life for diabetic patients, but when diet and exercise are not enough, medications may help diabetic patients control their blood sugar.



Bristol-Myers Squibb has generated an all-points attack on diabetes and related disorders. With a multi-pronged approach in metabolic research, the company will help ensure success in fully exploiting the opportunities provided in the post-genomics era.

In addition, we are investigating the emerging epidemic known as metabolic dysfunction. As more people live longer

and obesity becomes more prevalent, there are more people with borderline high levels of triglycerides, borderline high blood pressure and lower levels of HDL (high density lipoprotein) cholesterol. Each symptom on its own may not mark a person as diabetic, but the culmination of these risk factors may hasten obesity, atherosclerosis and osteoporosis and lead to greater disability or death.

Future treatments for diabetes and for many other diseases will likely consist of multiple drugs with different mechanisms rather than a single drug alone. Bristol-Myers Squibb is pursuing a multi-pronged, systematic approach to diabetes drug discovery and development. Many of our discovery programs have led to preclinical drug candidates.

Researchers around the world work tirelessly to discover and develop new medicines to fight metabolic diseases. This steadfast effort, coupled with the lessons we have learned and the vast potential of modern medicine, stands to benefit people everywhere.

Peter S. Ringrose, Ph. D.

*Chief Scientific Officer, Bristol-Myers Squibb and
President, Pharmaceutical Research Institute*

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Group European Manager:
Leonie Welss (4954)
European Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Silje Opstrup (4994)

Spain/ Portugal: Leonie Welss
(4954)

Production Manager:

Billie Franklin
To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Ashly de Leon

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakataamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejp.com
Japan Manager: Kate Cowan

naturejobs

Take control of your destiny

Information can provide an engine for change. The architects of a recently released US survey (see *Nature* **414**, 476; 2001) hope that data they have gathered from 32,000 graduate students will help to reform education — especially as the results indicate that one of the biggest shortcomings is the lack of information for students on careers outside academia.

The National Association of Graduate-Professional Students (NAGPS) based many of the survey questions on 'best practices' recommended by a variety of societies and government agencies. By collecting students' responses about how well their programmes enacted these recommendations, the association was able to grade and rank over 1,000 US graduate departments.

Lewis Siegel, vice-provost and dean of the graduate school at Duke University in Durham, North Carolina, says that the survey is "one of the most powerful tools I've seen in a decade" to get change in graduate education.

Naturejobs, in launching a new web page devoted to postdocs and graduate students, is providing another tool for students. This 'channel' (www.nature.com/naturejobs/channels/studentships/index.html) contains links to surveys such as the one conducted by the NAGPS, relevant government reports from around the world, and links to useful societies and organizational resources.

And, starting next month, *Naturejobs* will supplement this channel with a monthly feature about postdoc and studentship issues. As well as this new feature, the channel will also host *Naturejobs* articles about careers outside academia — just the kind of information many students who responded to the NAGPS survey said their academic advisers often failed to supply. In providing this resource, we hope to help students gain more control of their careers.

Paul Smaglik

Naturejobs editor



Contents

REGIONS

Australia's success stories in photonics and biotechnology p4

WWW.NATUREJOBS.COM

Career centre

Information on the scientific job market

FOCUS



SPOTLIGHT



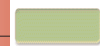
RECRUITMENT

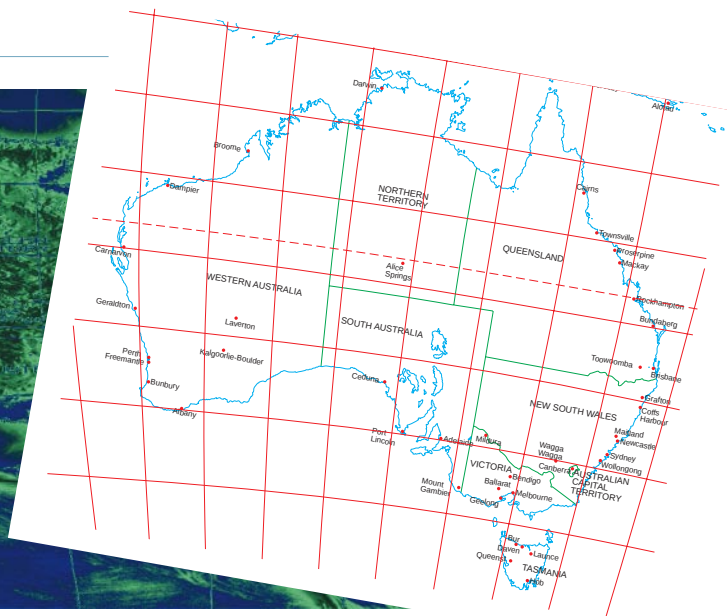
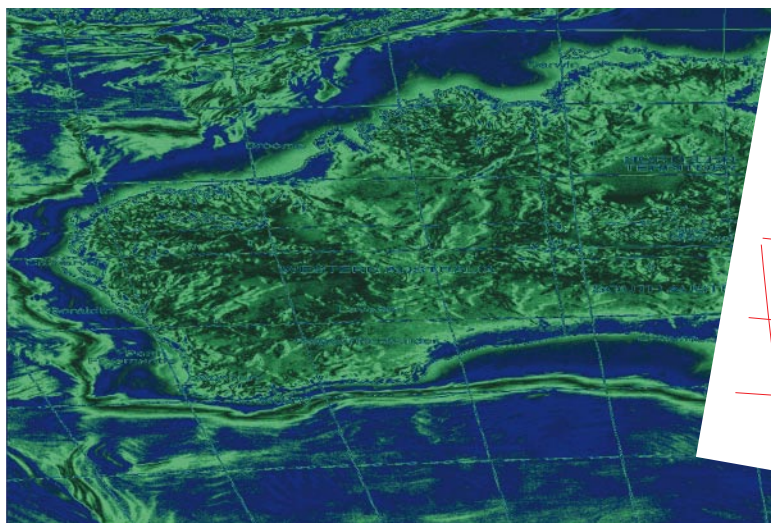


SCIENTIFIC ANNOUNCEMENTS



SCIENTIFIC EVENTS





Changing fortunes Australia

Young Australian scientists searching for work in their native country can face difficulties in both academic and industrial sectors. But for many, Australia's quality of life makes the search worthwhile.

In the academic sector, new PhDs looking for a faculty position might have a tough time because federal funding for universities has been decreasing since 1996. Vacancies are limited to replacing staff who have retired or left.

The alternative — a career in industry — offers only slightly better prospects. The private sector's low level of investment in research and development (R&D) has prevented biotechnology from booming to the extent that it has in the United States and parts of Europe. Since 1996, when the Conservative

government first cut tax incentives for in-house research, investment in industrial R&D has declined every year. And with the party's re-election last month, it is unlikely that the tide will turn in the near future.

But there are some bright spots. State and territory governments have recently started to compete to establish high-tech 'hubs' in their capital cities. And the federal government has begun to support more such targeted efforts.

In particular, photonics and biotechnology stand out as

success stories. Both sprouted from the cooperative research centres (CRCs), which the government established in 1991 to integrate research in universities and government labs with industry. The present government has committed A\$9.5 million (US\$4.9 million) for boosting infrastructure in photonics and A\$46.5 million for establishing one or more 'biotechnology centres of excellence'.

RAY OF LIGHT

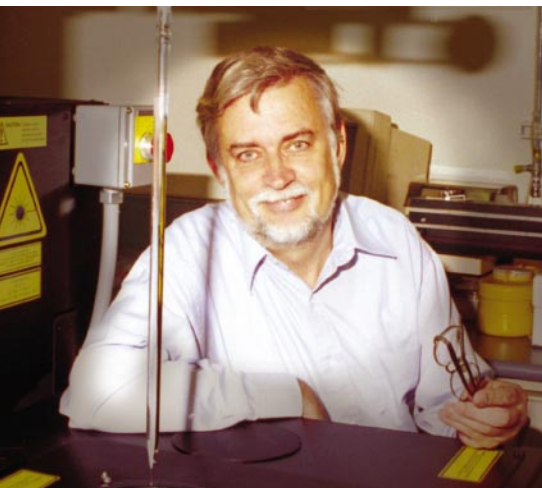
Photonics is the unsung hero of Australian R&D. Its rise began in 1986 when chemist Mark Sceats, physicist Ian Bassett and electrical engineer Tony Stokes had a few beers in the staff club at the University of Sydney and realized each was researching complementary aspects of optical fibres. The resulting Australian Photonics CRC is the largest of the 67 CRCs. It has 102 full-time-equivalent staff, of whom 75 are researchers, and 56 PhD students spread over five universities, 10 specialized spin-off companies and 29 business partners. Support from all sources totals A\$168 million over seven years.

Indx, the centre's manufacturing spin-off, was sold to US firm JDS Uniphase in 1997. Uniphase then transferred its Canadian work to an expanded plant in Sydney. It exports photonics components worth more than A\$100 million annually, all generated by research in the CRC and its family of companies.

Sceats, now chief executive of the photonics CRC, warns that industry must face the challenge of transforming itself to full automation. The balance of skills needed is moving from research in physics, chemistry and electrical engineering to mechanical, mechatronic and software engineering, he says, and

Light touch: Mark Sceats was instrumental in setting up Australia's photonics cooperative research centre.

P. POCKLEY





Glowing report: optical fibres are drawn from this holey polymeric fibre.

these are areas in which Australian universities have not been producing enough graduates.

The photonics CRC is also looking to double the number of researchers to 150 full-time staff equivalents and raise PhD numbers to 90 by 2005.

But there are still some questions about what skills photonics recruits will need. "It is quite good not to have been trained in photonics," says Maryanne Large, senior research fellow for the photonics CRC and based at the University of Sydney. Different skills and backgrounds lend themselves to more novel approaches, she says. Large is drawing on how three-dimensional structures in nature, such as the scales on a butterfly's wing, can act as diffraction gratings, a characteristic that

should prove valuable in polymeric optical fibres.

Although scientists pursuing photonics research in Australia generally make less money than they would if they moved to the United States, several prefer to stay (see 'Salary trade-offs', below). Photonics engineer Jen Bryce thinks of Sydney as "a fabulous place to live". Going back to her native Canada, she says, "would not improve my status". After a masters at the University of Sydney she joined Redfern Broadband Networks, one of the photonics CRC's spin-offs, where experienced

staff earn up to A\$110,000 plus share options.

Unlike photonics, biotech activity in Australia is distributed throughout the country. But disparities in housing prices mean that leaving a job in the cheaper, less populated cities to work in the more expensive, densely populated ones, can be difficult.

The Australian biotech industry is both younger and smaller than its counterpart in the United States. There are about 6,000 jobs and 280 companies in the sector as of 2001, says Peter Riddles, president of the industry association, AusBiotech. As in the United States, slightly higher pay compared with academic posts is making biotech a tempting alternative to doing a postdoc.

And increasing public-sector involvement in the industry is making some of the distinctions between the two sectors less relevant. For example, consortia are being formed in Melbourne and Brisbane to coordinate R&D in universities and medical research institutes, with an emphasis on industrial development.

In the past, such ventures have resulted in an increase in commercial offshoots. For example, CSL, a blood-products company based in Parkville, Victoria, was formed in 1994 when the Commonwealth Serum Laboratories were privatized. Last year, the company made A\$855 million and employed 1,300 in Australia and New Zealand, of whom 292 are researchers.

More recent efforts also show promise for job growth. For example, Proteome Systems started in Sydney two years ago with 14 staff, mostly from nearby Macquarie University. Now it has 70 staff, half of whom have PhDs.

A new arrangement established last month between the company and IBM could mean even more growth. The pair will collaborate across Proteome's research in protein analysis and drug discovery. Proteome's head and founder Keith Williams says this arrangement could result in a doubling of staff over the next year — especially in biosciences and information technology.

But not all biotech launches and spin-offs are so successful. The optimism expressed by participants in many start-up companies is tempered by sceptics who warn that many failures are inevitable in a relatively small country, and that the universities cannot deliver a sufficient base of skills in core sciences.

Biota offers the most cautionary story to date. The Melbourne-based company successfully developed the first all-Australian drug, the influenza treatment Relenza. Its share price rose to dizzying heights before crashing by nine-tenths when its initial global sales were not enough to deliver the predicted bonanza for investors. Biota has not failed — it is pressing ahead with production and R&D — but its wings have been trimmed.

Peter Pockley is Australasian correspondent for *Nature*.

Web links

Bio21 ♦ www.bio21.org

Institute for Molecular Bioscience ♦ www.imb.uq.edu.au



PHOTONICS CRC

Maryanne Large (left) believes it is helpful to approach photonics from a different background.

Salary trade-offs

Dax Kukulj and Robbie Charters of Redfern Polymer Optics (RPO) epitomize the diverse backgrounds of scientists moving into photonics in Australia.

After earning a PhD in polymer chemistry at the University of New South Wales, Kukulj worked on fabrics and hairsprays for chemical company Unilever in the United Kingdom. Bored by the routine work, he returned home in search of something more challenging. Now he is a project manager charged with developing optical fibres from polymers that can be processed at lower temperatures than glass fibres.

When Charters came

to Australia to do a postdoc seven years ago, he knew the laid-back atmosphere suited him. He has no desire to return to the more competitive United Kingdom. "Australia has an easier working environment," he says.

Both men acknowledge they've given up money for comfort, as starting pay of up to A\$70,000 (plus share options) in Australian high-tech is a fraction of what they could get in some US companies. But, as Canberra is a much cheaper place to live than many cities in Australia, they reckon that they're not losing out.

P.P.

In a flap: butterfly wings are under investigation at the photonics centre because their scales can behave as photonic crystals.

